Quantitative Methods for Analyzing Structure in Genomes, Self-Assembly, and Random Matrices

Quantitative Methods for Analyzing Structure in Genomes, Self-Assembly, and Random Matrices

A DISSERTATION PRESENTED
BY
MIRIAM HANNA HUNTLEY
TO
THE SCHOOL OF ENGINEERING AND APPLIED SCIENCES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
APPLIED MATHEMATICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
APRIL 2016

Thesis Advisors:

Professors Michael P. Brenner and Erez Lieberman Aiden

Author:

Miriam Hanna Huntley

Quantitative Methods for Analyzing Structure in Genomes, Self-Assembly, and Random Matrices

Abstract

This dissertation presents my graduate work analyzing biological structure. My research spans three different areas, which I discuss in turn. First I present my work studying how the genome folds. The three-dimensional structure of the genome inside of the nucleus is a matter of great biological importance, yet there are many questions about just how the genetic material is folded up. To probe this, we performed Hi-C experiments to create the highest resolution dataset (to date) of genome-wide contacts in the nucleus. Analysis of this data uncovered an array of fundamental structures in the folded genome. We discovered approximately 10,000 loops in the human genome, which each bring a pair of loci far apart along the DNA strand (up to millions of basepairs away) into close proximity. We found that contiguous stretches of DNA are segregated into self-associating contact domains. These domains are associated with distinct patterns of histone marks and segregate into six nuclear subcompartments. We found that these spatial structures are deeply connected to the regulation of the genome and cell function, suggesting that understanding and characterizing the 3D structure of the genome is crucial for a complete description of biology. Second, I present my work on self-assembly. Many biological structures are formed via 'bottom-up' assem-

bly, wherein a collection of subunits assemble into a complex arrangement. In this work we developed a theory which predicts the fundamental complexity limits for these types of systems. Using an information theory framework, we calculated the capacity, the maximum amount of information that can be encoded and decoded in systems of specific interactions, giving possible future directions for improvements in experimental realizations of self-assembly. Lastly, I present work examining the statistical structure of noisy data. Experimental datasets are a combination of signal and randomness, and data analysis algorithms, such as Principal Component Analysis (PCA), all seek to extract the signal. We used random matrix theory to demonstrate that even in situations where the dataset contains too much noise for PCA to be successful, the signal can be still be recovered with the use of prior information.

Contents

Acknowledgments

I have been very fortunate to receive support from many people over the course of my studies.

First and foremost, I would like to thank my two advisors, Erez Lieberman Aiden and Michael Brenner, who have both provided strong mentorship and guidance over the course of my studies and greatly influenced my career trajectory and scientific thinking.

I want to thank Erez for being an incredibly creative and enthusiastic advisor, and I feel very fortunate to have benefited from his guidance. His unbridled optimism, even in the face of setbacks, has many times over given me the motivation to continue moving forward. I am very grateful that he has always been willing to take as much time as necessary to discuss and improve any aspect of my work; he does not hesitate to dive into the smallest of details, going above and beyond what might be expected of any mentor. Erez also takes the responsibility of science communication very seriously, and he instills this sense of responsibility in his students, myself included. Erez's passion for science, brilliant creativity, and emblematic humor (i.e. terrible puns), have made working with him a unique experience from which I have learned a tremendous amount.

It has also been an absolute honor to work with Michael. Michael has taught me a great deal - the first lesson came my first year of graduate school in AM201, where he said "Don't be scared of messy equations." Michael has continued to teach me how to be courageous in the face of daunting calculations. His remarkable intuition for seeing

simple solutions to seemingly not-so-simple problems has been inspiring. I would like to thank Michael for giving me the freedom to pursue my interest in research topics that sometimes meandered quite a bit, often into dead ends, and for encouraging me that this was part of the scientific process. Michael has a reputation for his deep commitment and loyalty to his students, something I only fully understood once I started working with him and experienced this first hand: he has guided me through my studies and longer-term career decisions with unwavering support and kindness. I am grateful to have been one of the fortunate few who benefited from his mentorship.

I thank Michael Brenner, Erez Lieberman Aiden, and Radhika Nagpal for agreeing to be on my dissertation committee and for guiding me through my PhD. I also want to thank Radhika Nagpal, Ariel Amir and Shmuel Rubenstein for their support on my qualifying exam committee, and NSF for partially funding my studies through the GRFP.

If my work is to be judged by the company I've kept while working on it, I would count myself lucky. This thesis presents joint work that was done with many people, and I've gained a tremendous amount from these collaborations. I thank all of my coauthors for their dedication in making these papers happen. Chapter 2 of this thesis presents work I did in very close collaboration with Suhas Rao; I want to thank him for the many, many hours we spent working together, often without much sleep but always with amazing positivity, and for his stubborn commitment to getting science done. I also want to thank Neva Durand for her patience with my never-ending and often annoying requests for help, for setting an example for good coding practices, and for her always sage advice. I thank Elena Stamenova for teaching me how to do Hi-C experiments despite my inexperience. Chapter 4 presents joint work with Emily Darrow and Brian Chadwick, whose fascination with DXZ4 has been infectious and

who have both taught me much about the biology of the X chromosome. I also would like to thank Olga Dudchenko for her clear thinking and absolute reliability on this project. The fifth chapter of this thesis was joint work with Arvind Murugan, which was a wonderful collaboration; Arvind's unique way of approaching science has been a delight. The sixth chapter is joint work with Lucy Colwell, who, besides for her friendship and commitment to the project, has taught me a great deal about proteins and about navigating academia (both challenging subjects!). Many other academic collaborators deserve thanks for discussions, including all of the members of the Aiden Lab, the Brenner Group, and the Pierce Hall collective, especially Elizabeth Chen, Michael Chemama, Lihua Jin, Zorana Zeravcic, Michael Tikhonov, Sarah Kostinski, Muhammad Shamim, and Adrian Sanborn.

Life in graduate school would have been extremely dull without the deep friendships I've formed along the way, and though this thesis doesn't describe work with them per se, I was very influenced by the people that surrounded me during my time here. The Nashton community has been an absolute highlight of graduate school, and I am so glad that I was able to spend time with all of the (too many to name) wonderful roommates and friends who have been a part of it. I am grateful for the support over the years from friends old and new, during some of the toughest of times — looking at you, Melissa Lefkowitz, Noam Prywes, Dougal Maclaurin, Anders Sejr Hansen, Pepe Montiel Olea, Federico Villalpando, Sara Segal, and Libby Brockman. Thanks to the inestimable Noam Prywes for sharing his curiosity about the world with us by organizing the greatest journal club that ever was — ASAASA — where I could learn about topics as diverse as how pigeons navigate, how the internet works, and the tenets of Mormonism. And a very special thanks to the very special John Ingraham, who has enriched my life with his generosity, wit, and love and whose companionship I treasure immensely.

I cannot thank my family enough for their support over the years, but I can try. I want to thank my siblings Nena, Rafi, Ariel, and the later additions Andre and Michelle, for being the huge part of my life that they always have been, and for keeping me from taking myself too seriously. Time spent with them is a joy. Finally, I owe tremendous gratitude to my parents. They both have been my role models for how to live a true and meaningful life. They have also been my enablers in all my scholastic pursuits, even as a child, be it by buying me books as fast as I could read them or encouraging me when I wanted to teach myself programming. They have aided me and cheered me on through all of my endeavors, and I could not have done graduate school without their constant and unconditional support.

1

Introduction

This thesis describes my research analyzing structure in various biological systems. I will focus on the three-dimensional (3D) structure of genomes inside nuclei, on structures that can be built in self-assembly systems, and on structure in noisy datasets.

The cartoon model of the human genome paints a picture that the information content of our genomes is solely carried by the linear sequence of the four bases A,G,C,T (adenine, guanine, cytosine, and thymine). This is with good cause – our understanding of life has advanced a tremendous amount with the discovery of genes and the ability of the genome to transmit information from one generation to the next. Yet this picture is far from complete. Every cell in an organism has the same DNA, yet the morphology and function of cells from different tissues can be radically different. Where does the

information regarding which genes are expressed in a given cell get stored?

In the early 1980's, scientists discovered an unexpected feature of gene regulation: enhancers[9]. Enhancers are small genetic loci that act as switches that can 'turn on' the transcription of genes. The ability of these switches to turn on and off different genes provided a tantalizing avenue for explaining differentiation, or how the same genome can accomplish different functions in different circumstances. Intriguingly, though, it was found that enhancers could act as switches to genes that were not just adjacent to them along the DNA sequence, but also genes that were very far away, sometimes tens of thousands or millions of bases away[46]. This presented a new puzzle: how could enhancers affect genes that were so far away along the fiber? The proposed solution to this mystery was that the switches and the genes they were affecting were spatially co-located. Genomic loops, which bring together pairs of loci in the genome, are hypothesized to be the mechanism by which enhancers can act on genes which would otherwise be located far away. Yet to show that such features actually exist, experiments studying the 3D structure of the genome inside the nucleus were needed.

All of the DNA in a human cell, stretched end-to-end, would measure about 2 meters, and this long polymer is compacted inside a nucleus which has a size on the order of microns. We know now that the spatial structure of DNA inside the nucleus is far from random, and it is this non-randomness, and the effects of its organization, which many experiments have sought to untangle. At the nanometer scale, of course, we know that DNA forms a double-stranded helix[144]. In eukaryotes, this double helix is compacted by being wound around proteins called nucleosomes, forming the chromatin fiber[77,108]. This fiber is further folded inside the nucleus, yet the structure at length scales above hundreds of nanometers is not well defined. Early experiments revealed some general rules of genomic organization within the nucleus. For example, in the 1980's, elegant

work using DNA-damaging UV radiation discovered that chromosomes in the nucleus are not randomly mixed together like tangled spaghetti, but rather each segregate into self-associating chromosomal territories[35,34]. It was further discovered that chromatin which is gene-rich tends to be positioned towards the center of the nucleus, while gene-poor chromatin is clustered towards the nuclear periphery[33,78].

Despite these discoveries in the field of 3D genomics, the longstanding goal to uncover genomic loops was not easily resolved. Interrogation by conventional light microscopy is limited in resolution to the diffraction limit of 200 nm. The chromatin fiber is thought to have a diameter on the length scale of tens of nanometers, and so it is impossible to follow along its contour using optical techniques. This has led to the development of indirect experimental methods which can probe the 3D structure of the genome. One highly successful technique is that of proximity ligation[37,38]. This method allows the spatial structure in the genome to be measured not by optical technology but by sequencing technology. In these experiments, two genomic loci that are spatially adjacent to one another in the nucleus, even if they are far apart along the linear genome, are 'cut' from the remainder of the genome and then 'pasted' together into a chimeric piece of DNA. This chimera, once it is sequenced, reveals that the two loci were spatially collocated. Variants of this technology (3C[38], 4C[135], and 5C[42]) have been used to examine locus-locus interactions at specific locations. However, these methods could not catalog spatial interactions on the genome-wide scale.

Hi-C, invented in 2009[83], extended this technique to examine the 3D structure of the entire genome in a high-throughput manner. Hi-C experiments quantify how often any pair of loci in the genome come into contact. The higher the contact frequency of a pair, the more often they are found spatially collocated in the nucleus. Hi-C data is often represented in a matrix called a 'contact map', similar in spirit to adjacency matrices,

whose rows and columns index the loci in the genome and whose entries give the frequency with which these loci were in contact in an experiment, averaged over millions of cells. Studying contact maps reveals the statistical structure of the genome when it is folded up inside the nucleus. Hi-C has been successfully used to discover a wealth of spatial motifs of the folded genome. For example, it was found that the genome is partitioned into numerous domains that fall into two distinct compartments[41,72,83,134], separating 'active', transcribed chromatin from 'inactive' chromatin. Subsequent analyses suggested the presence of smaller domains, and have led to the important proposal that compartments are partitioned into condensed structures roughly one megabase in size, dubbed "topologically associated domains" (TADs)[41,105]. In principle, Hi-C could also be used to detect loops across the entire genome. To achieve this, however, extremely large data sets and rigorous computational methods were needed.

Chapter 2 of this thesis describes our studies of the 3D genome structure[119], in which we used Hi-C to generate the largest and highest-resolution contact map (to date). This allowed us to uncover an array of novel features of genome folding.

The most striking discovery we made was the existence of very well defined genomic loops, which connect two 'loop anchors', or genomic loci, hundreds of thousands to millions of basepairs apart. We identified close to 10,000 loops in the human genome, and characterized them in detail: they frequently link promoters and enhancers, they correlate with gene activation, and they show conservation across cell types and species. We found that the chromatin between the two loop anchors often forms a small contact domain, or a contiguous stretch of DNA about 200,000 basepairs long that strongly coassociates with itself, while being partitioned away from its immediate neighbors along the 1D genome. Analysis of these contact domains revealed that they are associated with distinct patterns of histone marks and segregate into six nuclear subcompartments,

each with distinct long-range contact patterns. This indicated a tight coupling between epigenetic information (given by histone modifications as well as other features), which is known to affect gene regulation, and spatial organization.

Upon further inspection of the loop anchors, we found that they almost without exception bind the DNA-binding protein CTCF. CTCF was previously thought to play a role in genomic looping[152,64], but its ubiquitous presence in the loops we discovered was surprising. Yet the surprise did not end there. When we investigated the DNA sequences at the exact locations that CTCF was binding at loop anchors, we discovered something quite shocking. CTCF is known to bind at a particular non-palindromic 'motif', or word in the genome, and so the two loop anchors of a loop comprise a pair of these motifs. In principle, these two motifs could be oriented in any which way with respect to each other. Strikingly, we found that in loops, the motifs are written predominantly in a convergent orientation, with the asymmetric motifs 'facing' one another. This gives a stringent and highly unanticipated rule for where loops can form in the human genome.

Finally, we used our dataset to explore whether the genome folds in a different manner in the two copies of each of the chromosomes. We discovered important instances of allele specific looping, which corroborated previous literature in the field regarding imprinting[79]. At large size scales, though, the two copies of the autosomal chromosomes fold in much same manner. On the other hand, the two copies of the X chromosome displayed radically different spatial organizational patterns. In human females, one of the X chromosomes is inactive, or turned off. We found that unlike in the active X chromosome, the inactive X chromosome splits into two massive domains (termed 'superdomains') and contains large loops (termed 'superloops') anchored at CTCF-binding repeats over tens of millions of basepairs away. This suggests that there is an

important link between the X spatial structure of the X chromosome and its inactivation.

Taken together, the results from this study represent a significant advance in our understanding of not only how the human genome is spatially organized, but also how this spatial organization affects biological function. It advocates for the view that a complete understanding of the genome is impossible with just the 1D information content of DNA, and must include a characterization of its 3D structure as well.

Prior to our work, there had been a number of studies, some using Hi-C and related proximity ligation techniques (such as 5C and ChIA-PET), which purported to have annotated the looping structure of all or a portion of the human genome. These studies suggested that there are near 1 million loops in the genome, in stark contrast to the 10,000 loops we annotate. In Chapter 3, I overview our work which shows that these previous loop annotations were almost entirely plagued by false positives. We believe that this was often due to neglecting to account for local spatial structure: if an algorithm annotates loops wherever the contact frequency of two loci is enriched above the genome-wide average, it will falsely annotate pairs of loci which are more frequently in contact for other reasons unrelated to looping. For example, two loci which are members of the same contact domain or the same subcompartment will be enriched for contact relative to the genome-wide average, yet would not be considered to be looping via a specific pairwise interaction. For this reason, it is crucial that an algorithm which annotates loops using contact data searches for pairs of loci that display focal enrichment, such that their contact frequency is large relative to the local neighborhood. To account for this, we developed our own loop-annotation algorithm, HiCCUPS.

In fact, the observation of novel high-resolution features in our Hi-C data required us to develop a whole suite of new scalable algorithms. Chapter 4 details a few of the new quantitative methods that I developed in order to study these structures, including: the

'Arrowhead' algorithm, a novel feature annotation algorithm that identified small blocks of co-associating DNA along the diagonal of contact matrices, which led to the conclusion that genomes are partitioned into small contact domains; analyses of epigenetic datasets in conjunction with Hi-C data, from which we inferred that contact domains exhibit distinct patterns of histone marks; a clustering algorithm that searched for similarity of patterns in the contact matrix, from which we concluded that the human genome segregates into at least six spatially distinct clusters or subcompartments; and Aggregate Peak Analysis (APA), which we use to validate lists of annotated loops by measuring the aggregate focal enrichment across all the putative loops.

Following this, in Chapter 5 I describe a followup study further investigating the structure of the inactive X chromosome. We had previously found that the inactive X was split into two superdomains at the locus DXZ4, and that a number of anchors, including DXZ4, formed a network of superloops, or long-range loops (connecting loci 5-70 million basepairs apart). To further investigate these features, I and my collaborators carried out Hi-C experiments in mouse and rhesus macaque cells, and found that both superdomains and superloops are present in these organisms, at the orthologous loci. This conservation across eutherian mammals suggests that these structural features may play an important functional role. We next sought to dive into studying the loci involved in creating superloops and superdomains. To do this we developed a variant of the Hi-C protocol, called COLA, which encouraged the ligation of three or more loci, instead of just two loci. Analyzing such data requires an entirely new pipeline: instead of 2D contact maps whose pixels represent contact frequency between any two loci, we created 3D contact tensors, whose voxels represented contact frequency between any three loci. With this in hand, we examined the triplet contact frequency of superloop anchors such as DXZ4 on the X chromosome, and found that they tend to co-locate

simultaneously, forming a spatial hub. Finally, we showed that deleting DXZ4 on the inactive X leads to the disappearance of superdomains and DXZ4-superloops, changes in compartmentalization patterns, and changes in the distribution of chromatin marks. Thus, we demonstrated that DXZ4 is essential for proper packaging of the inactive X.

Much of the work that I have done focuses on the discovery and characterization of 3D structures in the human genome. Yet such studies do not necessarily reveal the process by which such structures are assembled inside the cell to begin with. For example, how does a genomic loop ever form in the nucleus? Chapter 5 touches on this topic. It builds on previous work I collaborated on[123] in which we proposed the existence of an extrusion complex. This complex, we believe, lands on the chromatin and extrudes a loop by pulling the two strands of DNA through the complex. When the complex encounters a CTCF motif written in the correct orientation on a strand, that strand no longer can pass through the complex, while the other strand can continue extruding until a CTCF motif in the correct orientation is encountered on it as well. Thus the extrusion process will only halt when a convergently oriented pair of CTCF motifs is reached. The extrusion complex was hypothesized to explain the short-range loops seen throughout the genome, and is highly consistent with the data. In Chapter 5 we extended this model to explain the formation of superloops and superdomains seen only on the inactive X, thus conjecturing a unified mechanism for the assembly of genomic structures at very different size scales.

Loops in the human genome are only one of a staggering number of complex structures in a cell. Chapter 6 of this thesis moves away from studying the specifics of the 3D genome structure, and looks more generally at the process of self-assembling structures. Biology is regularly able to build incredible structures from the bottom-up, such as the ribosome, nuclear pores, or viral capsids. How biology is able to achieve such intricate

structures is still not fully understood. Synthetic experiments have attempted to duplicate biology's achievements in this area, but with limited success; despite rapid progress in this field, synthetically built self-assembled structures are still far less complex than the structures that have emerged over the course of evolution. To better understand why this is the case, my work in this area has focused on developing a deeper theoretical understanding of self-assembly.

Self-assembly (in both biological and synthetic contexts) is typically achieved by engineering binding specificities between the different types of particles, or 'species': each species must bind specifically to its target partner, and avoid binding with other off-target binding partners. Complex structures can be built up using these engineered specificities between species, but off-target interactions severely limit the number of species that can be used effectively. For example, suppose there is a system which creates particles whose surfaces like to stick to one another, but which can achieve binding specificity between 'on-target' pairs of particles by using shape complementarity – sculpting the surfaces of an on-target pair of particles such that they exactly fit together like a lock and key, making binding between them energetically favorable since this maximizes the surface area of contact. Simple combinatorics implies that the number of unique shapes that can be created in this system is exponential in the number of independent shape parameters that can be adjusted. Therefore, hierarchically building up more complex structures using these abundantly available species seems straightforward. In practice, however, the binding energies are never 100% specific; a particle might still bind with an off-target partner, because the two surfaces might have some amount of complementarity that yields a sufficient amount of binding energy. Because of this off-target binding, also known as crosstalk, a system which requires a high amount of specificity (on-target binding) will be restricted to use far fewer species than

the total number available.

In Chapter 6, I describe my work[65] studying what sets the limit for the number of interacting species a system can handle before the crosstalk overwhelms the specific interactions. To answer this, we reframed these systems in terms of information theory. Inspired by Shannon's Noisy Channel Coding Theorem, we treat desirable interactions between complementary on-target binding partners as the 'messages', and off-target binding events as 'errors'. This allows us to derive a 'capacity' for these systems, the maximal amount of mutual information that is encodable using engineered binding specificities. This provided us with a single framework for examining systems of specific interactions that used fundamentally different physics to achieve specificity. We compared various experimental methods, and found for example that systems which encoded specificity using shape complementarity have a higher capacity than systems that employ flat surface chemistry, e.g. via patterned areas of hydrophobic and hydrophilic patches. The reason for this difference in capacity is that the strength of off-target binding is linear in the size of the particles designed using flat chemistry specificities, while the off-target binding is only (at worst) sub-linear in the size of particles designed using shape complementarity. The mutual information framework also allowed us to examine how combining physics from two different systems, for example shape and surface chemistry, can lead to a nonlinear increase in the capacity. These and other insights give a more nuanced understanding of the limits of structures that can be built using self-assembly, and perhaps may aid in future improvements to synthetic experiments in self-assembly.

As the third topic in this thesis, in Chapter 7 I describe my work analyzing a different type of structure: structure in noisy datasets. In this age of increasing experimental capabilities, where millions or more variables can be measured at once, designing effec-

tive algorithms for analysis is crucial - one cannot just look at the data and intuit the underlying structure of the system. Hi-C experiments, for example, yield a genome-wide contact matrix of 3 million by 3 million entries. Discovering the biological structure from such a dataset first requires discovering the statistical structure of the data.

One of the most widely used techniques to uncover statistical structure in large datasets is Principal Component Analysis (PCA), which yields principal components, or signal vectors which annotate the most significant sources of covariance in the data. However, in sample-limited scenarios where many variables are measured by a limited number of samples, these datasets can be noisy, and algorithms such as PCA can be quite limited in their ability to extract signal. In the analysis of some biological experiments which operate in this sample-limited regime, it has become increasingly popular for researchers to perform 'pragmatic' adjustments to the data before PCA is carried out, wherein the data matrix is first transformed in some fashion to incorporate implicit prior knowledge about the structure of the data. These methods account for various types of prior information about the system, for example knowledge that certain samples are correlated or corrupted, or partial knowledge of which variables correlated. The legitimacy of these pragmatic PCA methods is largely determined ex post facto, with the quality of the results used to justify the validity of the procedure.

In this project, detailed in Chapter 7, I set out to understand the justifications and limits of pragmatic PCA techniques. To do this I used tools from Random Matrix Theory (RMT), which give quantitative predictions as to how random noise can affect signal detection in PCA. With RMT one can compute the expected overlap of the principal components with the true signal, with the surprising result that a signal-to-noise ratio which does not exceed a particular signal detectability threshold will never be recoverable by PCA. Extending these ideas, I studied how linear transformations of the

data which incorporate prior information can improve the signal detectability. Broadly speaking, our results are intuitive: employing a well-informed prior with true knowledge of the system can abnegate some of the effects of noise and markedly improve signal detection by PCA, while the use of an incorrect prior can completely destroy the signal in the principal components. We use these tools to examine examples of pragmatic PCA from the recent literature, including Statistical Coupling Analysis and outlier deletion, and quantify the expected theoretical improvements in signal detectability. Thus, our work offers a generally applicable tool for understanding why many of the often unjustified techniques proposed in recent literature can be so efficacious at uncovering structure in noisy data.

Finally, in Chapter 8 I conclude with thoughts regarding the outlook of these fields. We have made significant progress in understanding the biological structure in the genome, in self-assembly, and in noisy datasets. And yet, far more remains to be uncovered.

2

A three-dimensional map of the human genome at kilobase resolution reveals principles of chromatin looping

2.1 Introduction

The spatial organization of the human genome is known to play an important role in the transcriptional control of genes[16,34,133]. Yet important questions remain, like how distal regulatory elements, such as enhancers, affect promoters, and how insulators can abrogate these effects[9,17,49]. Both phenomena are thought to involve the formation of

protein-mediated "loops" that bring pairs of genomic sites that lie far apart along the linear genome into proximity[127].

Various methods have emerged to assess the three-dimensional architecture of the nucleus. In one seminal study, the binding of a protein to sites at opposite ends of a restriction fragment created a loop, which was detectable because it promoted the formation of DNA circles in the presence of ligase. Removal of the protein or either of its binding sites disrupted the loop, eliminating this "cyclization enhancement."[97]. Subsequent adaptations of cyclization enhancement made it possible to analyze chromatin folding in vivo, including nuclear ligation assay[37] and chromosome conformation capture[38], which analyze contacts made by a single locus, extensions such as 5C for examining several loci simultaneously[42], and methods such as ChIA-PET for examining all loci bound by a specific protein[48].

To interrogate all loci at once, we developed Hi-C, which combines DNA proximity ligation with high-throughput sequencing in a genome-wide fashion[83]. We used Hi-C to demonstrate that the genome is partitioned into numerous domains that fall into two distinct compartments[41,72,83,134]. Subsequent analyses have suggested the presence of smaller domains, and have led to the important proposal that compartments are partitioned into condensed structures roughly one megabase in size, dubbed "topologically associated domains" (TADs)[41,105]. In principle, Hi-C could also be used to detect loops across the entire genome. To achieve this, however, extremely large data sets and rigorous computational methods are needed. Recent efforts have suggested that this is an increasingly plausible goal[69,134].

Here, we report the results of an effort to comprehensively map chromatin contacts genome-wide, using in situ Hi-C, in which DNA-DNA proximity ligation is performed in intact nuclei. The protocol facilitates the generation of much denser Hi-C maps.

The maps reported here comprise over 5 terabases of sequence data recording over 15 billion distinct contacts, an order of magnitude larger than all published Hi-C datasets combined. Using these maps, we are able to clearly discern domain structure, compartmentalization, and thousands of chromatin loops. In addition to haploid maps, we were also able to create diploid maps analyzing each chromosomal homolog separately. The maps provide a picture of genomic architecture with resolution down to 1 kilobase.

2.2 Results

2.2.1 In situ Hi-C methodology and maps

Our in situ Hi-C protocol combines our original Hi-C protocol (here called dilution Hi-C) with nuclear ligation assay[37], in which DNA is digested using a restriction enzyme, DNA-DNA proximity ligation is performed in intact nuclei, and the resulting ligation junctions are quantified. Our in situ Hi-C protocol involves cross-linking cells with formaldehyde; permeabilizing them with nuclei intact; digesting DNA with a suitable 4-cutter restriction enzyme (such as MboI); filling the 5'-overhangs while incorporating a biotinylated nucleotide; ligating the resulting blunt-end fragments; shearing the DNA; capturing the biotinylated ligation junctions with streptavidin beads; and analyzing the resulting fragments with paired-end sequencing (Figure 2.1A). This protocol resembles a recently published single-cell Hi-C protocol[101], which also performed DNA-DNA proximity ligation inside nuclei to study nuclear architecture in individual cells. Our updated protocol has three major advantages over dilution Hi-C. First, in situ ligation reduces the frequency of spurious contacts due to random ligation in dilute solution, as evidenced by a lower frequency of junctions between mitochondrial and nuclear DNA in the captured fragments, and by the higher frequency of random ligations observed

when the supernatant is sequenced. This is consistent with a recent study showing that ligation junctions formed in solution are far less meaningful[50]. Second, the protocol is faster, requiring three days instead of seven. Third, it enables higher resolution and more efficient cutting of chromatinized DNA, for instance, through the use of a 4-cutter rather than a 6-cutter.

A Hi-C map is a list of DNA-DNA contacts produced by a Hi-C experiment. By partitioning the linear genome into "loci" of fixed size (e.g., bins of 1 Mb or 1 kb), the Hi-C map can be represented as a "contact matrix" M, where the entry $M_{i,j}$ is the number of contacts observed between locus L_i and locus L_j. (A "contact" is a read pair that remains after we exclude reads that are duplicates, that correspond to unligated fragments, or that that do not align uniquely to the genome.) The contact matrix can be visualized as a heatmap, whose entries we call "pixels". An "interval" refers to a set of consecutive loci; the contacts between two intervals thus form a "rectangle" or "square" in the contact matrix. We define the "matrix resolution" of a Hi-C map as the locus size used to construct a particular contact matrix and the "map resolution" as the smallest locus size such that 80% of loci have at least 1000 contacts. The map resolution is meant to reflect the finest scale at which one can reliably discern local features.

2.2.2 Contact maps spanning 9 cell lines containing over 15 billion contacts.

We constructed in situ Hi-C maps of 9 cell lines in human and mouse. Whereas our original Hi-C experiments had a map resolution of 1 Mb, these maps have a resolution of 1 kb or 5 kb. Our largest map, in human GM12878 B-lymphoblastoid cells, contains 4.9 billion pairwise contacts and has a map resolution of 950 bp ("kilobase resolution"). We also generated eight in situ Hi-C maps at 5 kb resolution, using cell lines representing all human germ layers (IMR90, HMEC, NHEK, K562, HUVEC, HeLa, and KBM7) as well as mouse B-lymphoblasts (CH12-LX). Each map contains between 395M and 1.1B contacts.

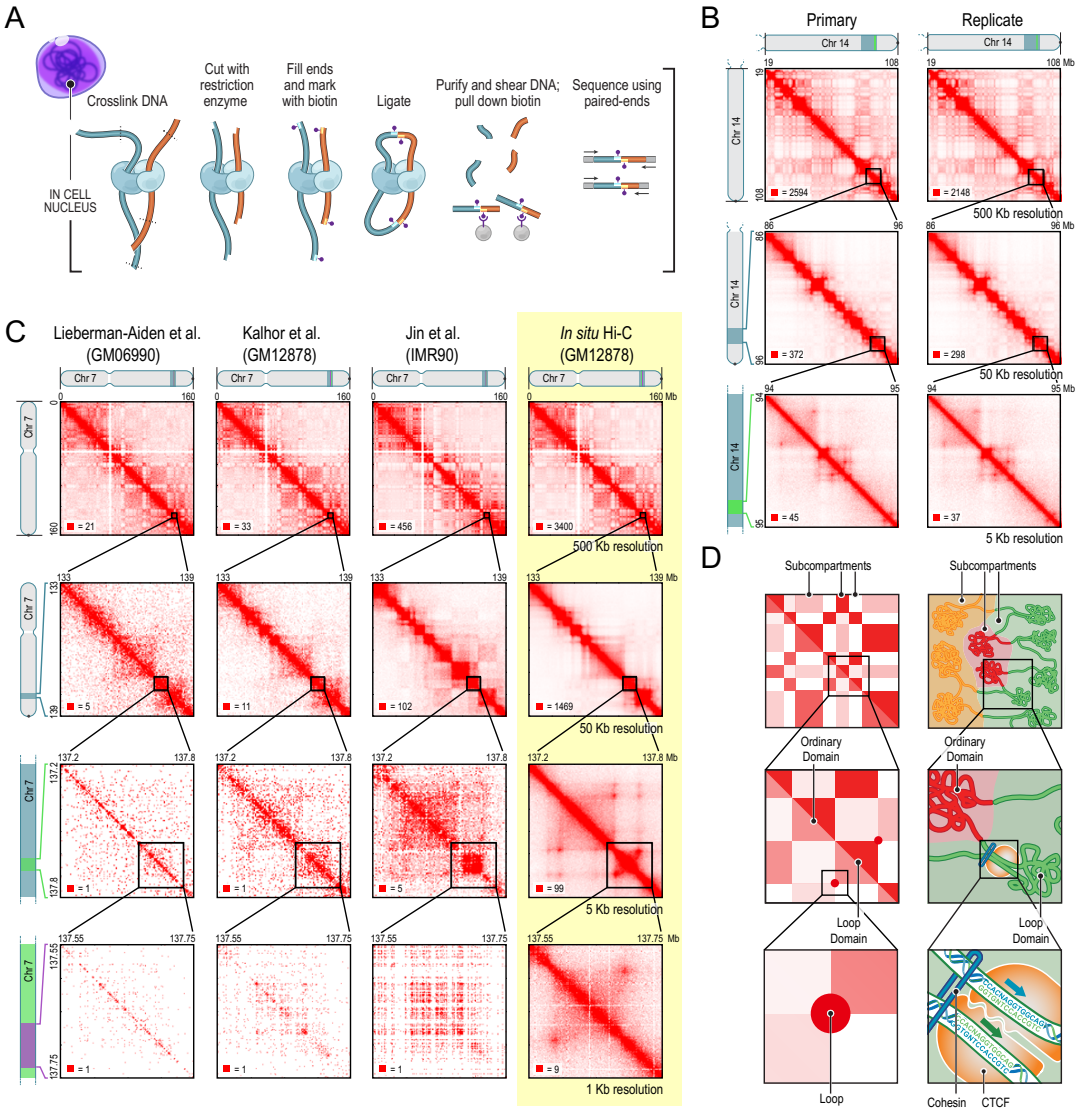
When we used our original dilution Hi-C protocol to generate maps of GM12878, IMR90, HMEC, NHEK, HUVEC, and CH12-LX, we found that, as expected, in situ Hi-C maps were superior at high resolutions, but closely resembled dilution Hi-C at lower resolutions. For instance, our dilution map of GM12878 (3.2 billion contacts) correlated highly with our in situ map at 500, 50, and 25 kb resolutions ($R>0.96,0.90,0.87$ respectively).

We also performed 112 supplementary Hi-C experiments using three different protocols (in situ Hi-C, dilution Hi-C, and Tethered Conformation Capture) while varying a wide array of conditions such as extent of crosslinking, restriction enzyme, ligation volume/time, and biotinylated nucleotide. These include several in situ Hi-C experiments in which the formaldehyde crosslinking step was omitted, which demonstrate that the structural features we observe cannot be due to the crosslinking procedure. In total, 201 independent Hi-C experiments were successfully performed.

To account for non-uniformities in coverage due to the number of restriction sites

Figure 2.1 *(following page)*: We Used In Situ Hi-C to Map over 15 Billion Chromatin Contacts across Nine Cell Types in Human and Mouse, Achieving 1 kb Resolution in Human Lymphoblastoid Cells. (A) During in situ Hi-C, DNA-DNA proximity ligation is performed in intact nuclei. (B) Contact matrices from chromosome 14: the whole chromosome, at 500 kb resolution (top); 86-96 Mb/50 kb resolution (middle); 94-95 Mb/5 kb resolution (bottom). Left: GM12878, primary experiment; Right: biological replicate. The 1D regions corresponding to a contact matrix are indicated in the diagrams above and at left. The intensity of each pixel represents the normalized number of contacts between a pair of loci. Maximum intensity is indicated in the lower left of each panel. (C) We compare our map of chromosome 7 in GM12878 (last column) to earlier Hi-C maps: Lieberman-Aiden et al. (2009), Kalhor et al. (2012), and Jin et al. (2013). (D) Overview of features revealed by our Hi-C maps. Top: the long- range contact pattern of a locus (left) indicates its nuclear neighborhood. We detect at least six subcompartments, each bearing a distinctive pattern of epigenetic features. Middle: squares of enhanced contact frequency along the diagonal (left) indicate the presence of small domains of condensed chromatin, whose median length is 185 kb (right). Bottom: peaks in the contact map (left) indicate the presence of loops (right). These loops tend to lie at domain boundaries and bind CTCF in a convergent orientation.

Figure 2.1: (continued)



at a locus or the accessibility of those sites to cutting[83,149] we use a matrix-balancing algorithm due to P. Knight and D. Ruiz[76].

Adequate tools for visualization of these large data sets are essential. We have therefore created the Juicebox visualization system that enables users to explore contact matrices, zoom in and out, compare Hi-C matrices to 1D tracks, superimpose all features reported in this paper onto the data, and contrast different Hi-C maps. All contact data and feature sets reported here can be explored interactively via Juicebox at: http://www.aidenlab.org/juicebox.

2.2.3 The genome is partitioned into small domains whose median length is 185 kb.

We began by probing the three-dimensional partitioning of the genome. In our earlier experiments at 1 Mb map resolution[83], we saw large squares of enhanced contact frequency tiling the diagonal of the contact matrices. These squares partitioned the genome into 5-20 Mb intervals, which we here call "megadomains."

We also found that individual 1 Mb loci could be assigned to one of two long-range contact patterns, which we called Compartments A and B, with loci in the same compartment showing more frequent interaction. Megadomains — and the associated squares along the diagonal — arise when all of the 1 Mb loci in an interval exhibit the same genome-wide contact pattern. Compartment A is highly enriched for open chromatin; Compartment B is enriched for closed chromatin[72,83,134].

In our new, higher resolution maps (200- to 1000-fold more contacts), we observe many small squares of enhanced contact frequency that tile the diagonal of each contact matrix (Figure 2.2A). We used the Arrowhead algorithm (see Experimental Procedures) to annotate these contact domains genome-wide. The observed domains range in size

from 40 kb to 3 Mb (median size 185 kb). As with megadomains, there is an abrupt drop in contact frequency (33%) for pairs of loci on opposite sides of the domain boundary. Contact domains are often preserved across cell types.

The presence of smaller domains in Hi-C maps is consistent with several other recent studies[41,105,134]. We explore the relationship between the domains we annotate and those annotated in prior studies in the Discussion.

2.2.4 Contact domains exhibit consistent histone marks, whose changes are associated with changes in long-range contact pattern

Loci within a contact domain show correlated histone modifications for eight different factors (H3K36me3, H3K27me3, H3K4me1, H3K4me2, H3K4me3, H3K9me3, H3K79me2, and H4K20me1) based on data from the ENCODE project in GM12878 cells[30]. By contrast, loci at comparable distance but residing in different domains showed much less correlation in chromatin state (Figure 2.2B, Extended Experimental Procedures). Strikingly, changes in a domain's chromatin state are often accompanied by changes in the long-range contact pattern of domain loci (i.e., the pattern of contacts between loci in the domain and other loci genome-wide), indicating that changes in chromatin pattern are accompanied by shifts in a domain's nuclear neighborhood (Figure 2.2C, Extended Experimental Procedures). This observation is consistent with microscopy studies associating changes in gene expression with changes in nuclear localization[44].

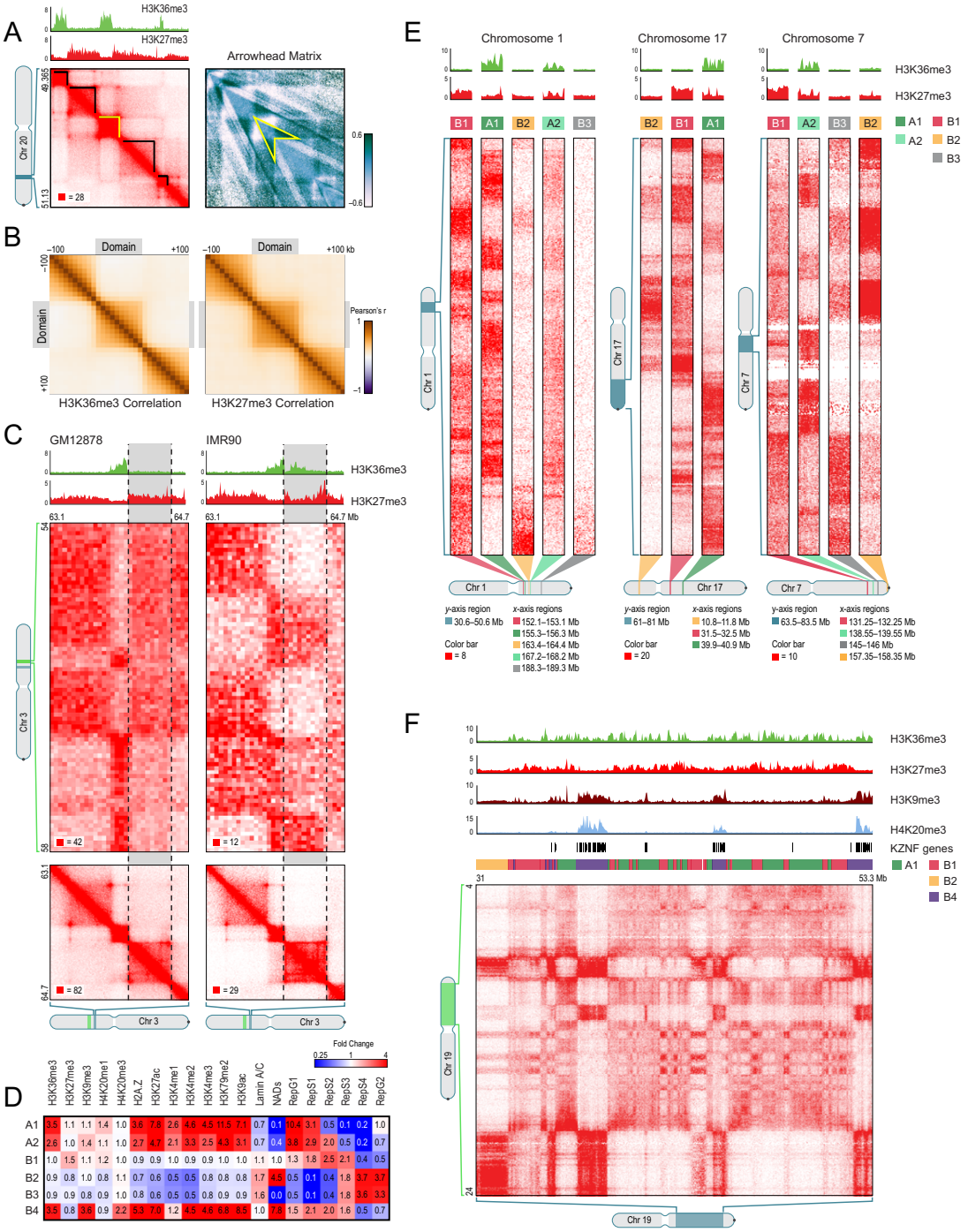
2.2.5 There are at least six nuclear subcompartments with distinct patterns of histone modifications

Next, we partitioned loci into categories based on long-range contact patterns alone, using four independent approaches: manual annotation, and three unsupervised clustering algorithms (HMM, K-means, Hierarchical). All gave similar results. We then investigated the biological meaning of these categories.

When we analyzed the data at low matrix resolution (1 Mb), we reproduced our earlier finding of two compartments (A and B). At high resolution (25 kb), we found evidence for at least five "subcompartments" defined by their long-range interaction patterns, both within and between chromosomes. These findings expand on earlier

Figure 2.2 *(following page)*: The Genome Is Partitioned into Contact Domains That Segregate into Nuclear Subcompartments Corresponding to Different Patterns of Histone Modifications. (A) We annotate thousands of domains across the genome (left, black highlight). To do so, we define an arrowhead matrix A (right) such that $A_{i,i+d} = (M^*_{i,i-d} - M^*_{i,i+d})/(M^*_{i,i-d} + M^*_{i,i+d})$, where M^* is the normalized contact matrix. This transformation replaces domains with an arrowhead-shaped motif pointing toward the domain's upper-left corner (example in yellow); we identify these arrowheads using dynamic programming. See Experimental Procedures. (B) Pearson correlation matrices of the histone mark signal between pairs of loci inside and within 100 kb of a domain. Left: H3K36me3; Right: H3K27me3. (C) Conserved contact domains on chromosome 3 in GM12878 (left) and IMR90 (right). In GM12878, the highlighted domain (gray) is enriched for H3K27me3 and depleted for H3K36me3. In IMR90, the situation is reversed. Marks at flanking domains are the same in both: the domain to the left is enriched for H3K36me3 and the domain to the right is enriched for H3K27me3. The flanking domains have long-range contact patterns that differ from one another and are preserved in both cell types. In IMR90, the central highlighted domain is marked by H3K36me3 and its long-range contact pattern matches the similarly-marked domain on the left. In GM12878, it is decorated with H3K27me3, and the long-range pattern switches, matching the similarly-marked domain to the right. Diagonal submatrices, 10 kb resolution; long-range interaction matrices, 50 kb resolution. (D) Each of the six long-range contact patterns we observe exhibits a distinct epigenetic profile. Each subcompartment also has a visually distinctive contact pattern. (E) Each example shows part of the long-range contact patterns for several nearby genomic intervals lying in different compartments. (F) A large contiguous region on chromosome 19 contains intervals in subcompartments A1, B1, B2, and B4.

Figure 2.2: (continued)



reports suggesting three compartments in human cells[149]. We found that the median length of an interval lying completely within a subcompartment is 300 kb. Although the subcompartments are defined solely based on their Hi-C interaction patterns, they exhibit distinct genomic and epigenomic content.

Two of the five interaction patterns are correlated with loci in compartment A. We label the loci exhibiting these patterns as belonging to subcompartments A1 and A2. Both A1 and A2 are gene dense, have highly expressed genes, harbor activating chromatin marks such as H3K36me3, H3K79me2, H3K27ac and H3K4me1 and are depleted at the nuclear lamina and at nucleolus associated domains (NADs). (Figure 2.2D,E,) While both A1 and A2 exhibit early replication times, A1 finishes replicating at the beginning of S-phase, whereas A2 continues replicating into the middle of S-phase. A2 is more strongly associated with the presence of H3K9me3 than A1, has lower GC content, and contains longer genes (2.4-fold).

The other three interaction patterns (labeled B1, B2, and B3) are correlated with loci in compartment B, and show very different properties. Subcompartment B1 correlates positively with H3K27me3 and negatively with H3K36me3, suggestive of facultative heterochromatin (Figure 2.2D,E). Replication of this subcompartment peaks during the middle of S-phase. Subcompartments B2 and B3 tend to lack all of the above-noted marks, and do not replicate until the end of S-phase (See Figure 2.2D). Subcompartment B2 includes 62% of pericentromeric heterochromatin (3.8-fold enrichment) and is enriched at the nuclear lamina (1.8-fold) and at NADs (4.6-fold). Subcompartment B3 is enriched at the nuclear lamina (1.6-fold), but strongly depleted at NADs (76-fold).

Upon closer visual examination, we noticed the presence of a sixth pattern on chromosome 19 (Figure 2.2F). Our genome-wide clustering algorithm missed this pattern because it spans only 11 Mb, or 0.3% of the genome. When we repeated the algorithm

on chromosome 19 alone, the additional pattern was detected. Because this sixth pattern correlates with the Compartment B pattern, we labeled it B4. Subcompartment B4 comprises a handful of regions, each of which contains many KRAB-ZNF superfamily genes. (B4 contains 130 of the 278 KRAB-ZNF genes in the genome, a 65-fold enrichment). As noted in previous studies[56,142], these regions exhibit a highly distinctive chromatin pattern, with strong enrichment for both activating chromatin marks, such as H3K36me3, and heterochromatin-associated marks, such as H3K9me3 and H4K20me3.

2.2.6 Approximately 10,000 peaks mark the position of chromatin loops

We next sought to identify the positions of chromatin loops by using an algorithm to search for pairs of loci that show significantly closer proximity with one another than with the loci lying between them (Figure 2.3A). Such pairs correspond to pixels with higher contact frequency than typical pixels in their neighborhood. We refer to these pixels as "peaks" in the Hi-C contact matrix, and to the corresponding pair of loci as "peak loci". Peaks reflect the presence of chromatin loops, with the peak loci being the anchor points of the chromatin loop. (Because contact frequencies vary across the genome, we define peak pixels relative to the local background. We note that some papers[69,124] have sought to define peaks relative to a genome-wide average. This choice is problematic because, for example, many pixels within a domain may be reported as peaks despite showing no locally distinctive proximity; see Discussion.)

Our algorithm detected 9,448 peaks in the in situ Hi-C map for GM12878 at 5 kb matrix resolution. These peaks are associated with a total of 12,903 distinct peak loci (some peak loci are associated with more than one peak). The vast majority of peaks (98%) reflected loops between loci that are less than 2 Mb apart.

These findings were reproducible across all of our high-resolution Hi-C maps. Examining the primary and replicate maps separately, we found 8,054 peaks in the former and 7,484 peaks in the latter, with 5,403 in both lists; see figures 2.3A, and 2.3B. The differences were almost always the result of our conservative peak-calling criteria. We also called peaks using our GM12878 dilution Hi-C experiment. Because the map is sparser and thus noisier, we called only 3,073 peaks. Nonetheless, 65% of these peaks were also present in the list of peaks from our in situ Hi-C dataset, again reflecting inter-replicate reproducibility.

To independently confirm that peak loci are closer than neighboring locus pairs, we performed 3D-FISH[12] on 4 loops. In each case, we compared two peak loci, L_1 and L_2, with a control locus, L_3, that lies an equal distance away from L_2 but on the opposite side (Figure 2.3C). In all cases, the distance between L_1 and L_2 was consistently shorter than the distance between L_2 and L_3.

We also confirmed that our list of peaks was consistent with previously published Hi-C maps. Although earlier maps contained too few contacts to reliably call individual peaks, we developed a method called Aggregate Peak Analysis (APA) that compares the aggregate enrichment of our peak set in these low-resolution maps to the enrichment seen when our peak set is translated in any direction (Experimental Procedures). APA showed strong consistency between our loop calls and all six previously published Hi-C datasets for lymphoblastoid cell lines[72,83] (Figure 2.3D).

Finally, we demonstrated that the peaks observed were robust to particular protocol conditions by performing APA on our GM12878 dilution Hi-C map, and on our 112 supplemental Hi-C experiments exploring a wide range of protocol variants. Enrichment was seen in every experiment. Notably, these include five experiments (HIC043-HIC047) in which the Hi-C protocol was performed without crosslinking, demonstrating

that the peaks observed in our experiments cannot be byproducts of the formaldehyde-crosslinking procedure.

2.2.7 Conservation of peaks among human cell lines and across evolution

We also identified peaks in the other six human cell lines. Because these maps contain fewer contacts, sensitivity is reduced, and fewer peaks are observed (ranging from 2634 to 8040). APA confirmed that these peak calls were consistent with the dilution Hi-C maps reported here (in IMR90, HMEC, HUVEC, and NHEK), as well as with all previously published Hi-C maps in these cell types[41,69,83].

We found that peaks were often conserved across cell types (Figure 2.4A): between 55% and 75% of the peaks found in any given cell type were also found in GM12878.

Next, we compared peaks across species. In CH12-LX mouse B-lymphoblasts, we identified 2927 high-confidence contact domains and 3331 peaks. When we examined orthologous regions in GM12878, we found that 50% of peaks and 45% of domains called in mouse were also called in humans. This suggests substantial conservation of three-dimensional genome structure across the mammals (Figure 2.4B-E).

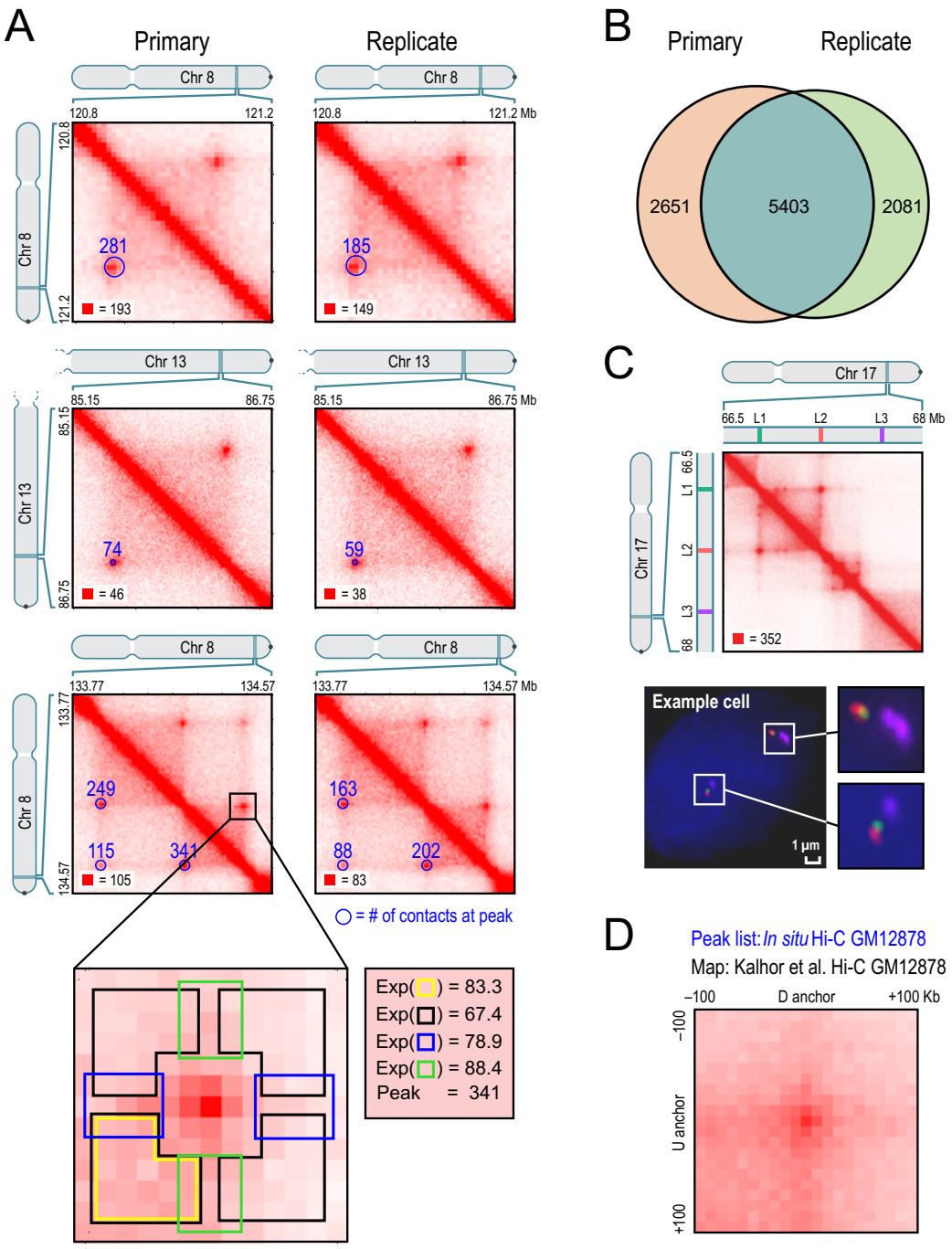
2.2.8 Loops anchored at a promoter are associated with enhancers and increased gene activation

Various lines of evidence indicate that many of the observed loops are associated with gene regulation.

First, our peaks frequently have a known promoter at one peak locus (as annotated by ENCODE's ChromHMM[62]), and a known enhancer at the other (Figure 2.5A). For instance, 2854 of the 9448 peaks in our GM12878 map bring together known promoters and known enhancers (30%, vs. 7% expected by chance). The peaks include classic promoter-enhancer loops, such as at MYC (chr8:128.35-128.75 Mb, in HMEC) and

Figure 2.3 *(following page)*: We Identify Thousands of Chromatin Loops Genome-wide Using a Local Background Model. (A) We identify peaks by detecting pixels that are enriched with respect to four local neighborhoods (blowout): horizontal (blue), vertical (green), lower-left (yellow), and donut (black). These "peak" pixels indicate the presence of a loop, and are marked with blue circles (radius = 20 kb) in the lower-left of each heatmap. The number of raw contacts at each peak is indicated. Left: primary GM12878 map; Right: replicate; annotations are completely independent. All contact matrices in these figures are 10 kb resolution unless noted. (B) Overlap between replicates. (C) Top: location of 3D-FISH probes. Bottom: example cell. (D) APA plot shows the aggregate signal from the 9,4948 GM12878 loops we report by summing submatrices surrounding each peak in a low-resolution GM12878 Hi-C map due to Kalhor et al. (2012). Although individual peaks cannot be discerned in the Kalhor et al. data (which contains 42M contacts), the peak at the center of the APA plot indicates that the aggregate signal from our peak set as a whole can be clearly discerned using their dataset.

Figure 2.3: (continued)



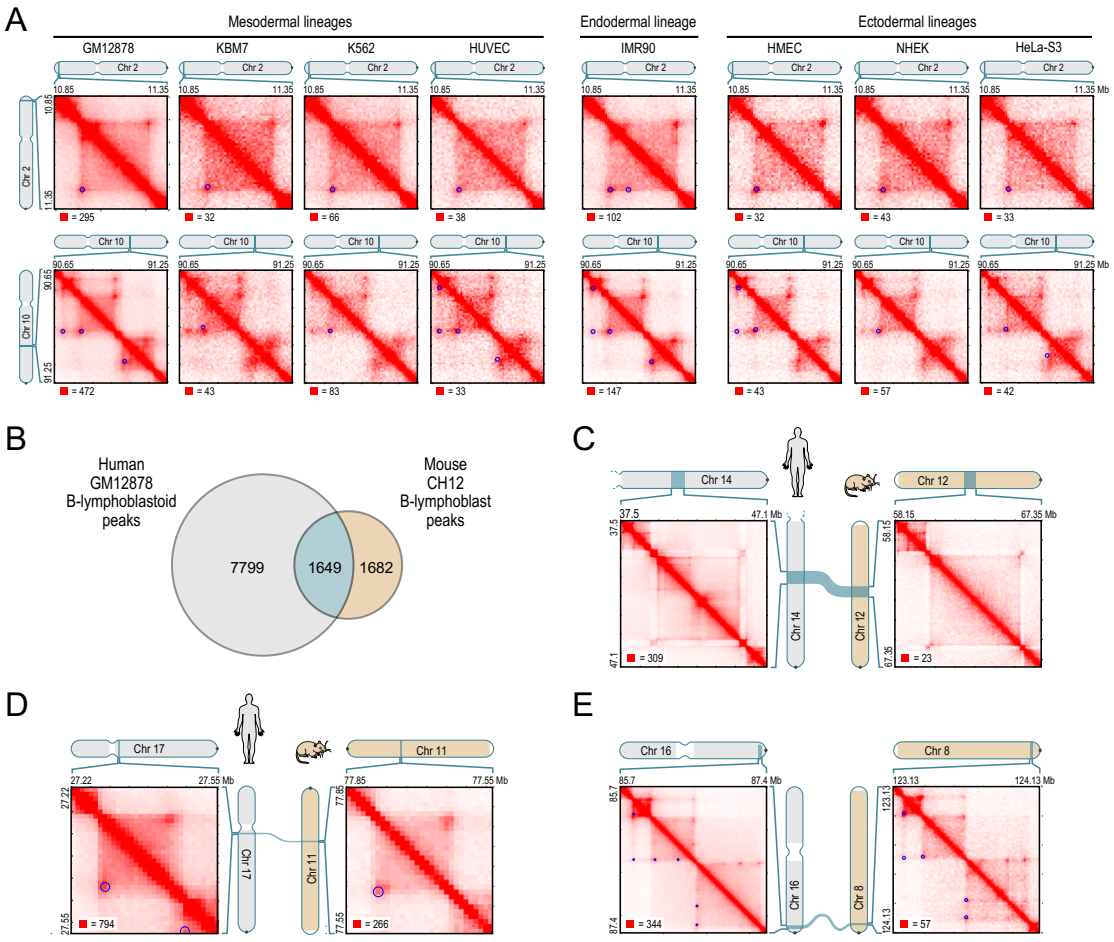


Figure 2.4: Loops Are Often Preserved across Cell Types and from Human to Mouse. (A) Examples of peak and domain preservation across cell types. Annotated peaks are circled in blue. All annotations are completely independent. (B) Of the 3,331 loops we annotate in mouse CH12-LX, 1,649 (50%) are orthologous to loops in human GM12878. (C-E) Conservation of 3D structure in synteny blocks.

alpha-globin (chr16:0.15-0.22 Mb, in K562). Second, genes whose promoters are associated with a loop are much more highly expressed than genes whose promoters are not associated with a loop (6-fold).

Third, the presence of cell type-specific peaks is associated with changes in expression. When we examined RNA-Seq data produced by ENCODE, we found that the appearance of a loop in a cell type was frequently accompanied by the activation of a gene whose promoter overlapped one of the peak loci. For example, a cell-type specific loop is anchored at the promoter of the gene encoding L-selectin (SELL), which is expressed in GM12878 (where the loop is present) but not in IMR90 (where the loop is absent, Figure 2.5B). Genome-wide, we observed 557 loops in GM12878 that were clearly absent in IMR90. The corresponding peak loci overlapped the promoters of 43 genes that were markedly upregulated (>50-fold) in GM12878, but of only 1 gene that was markedly upregulated in IMR90. Conversely, we found 510 loops in IMR90 that were clearly absent in GM12878. The corresponding peak loci overlapped the promoters of 94 genes that were markedly upregulated in IMR90, but of only 3 genes that were markedly upregulated in GM12878. When we compared GM12878 to the five other human cell types for which ENCODE RNA-Seq data was available, the results were very similar (Figure 2.5C).

Occasionally, gene activation is accompanied by the emergence of a cell-type-specific network of peaks. Figure 2.5D illustrates the case of ADAMTS1, which encodes a protein involved in fibroblast migration. The gene is expressed in IMR90, where its promoter is involved in six loops. In GM12878, it is not expressed, and the promoter is involved in only two loops. Many of the IMR90 peak loci form transitive peaks with one another (see discussion of "transitivity" below), suggesting that the ADAMTS1 promoter and the six distal sites may all be located at a single spatial hub.

These observations are consistent with the classic model in which looping between a promoter and enhancer activates a target gene[2,6,139].

2.2.9 Loops frequently demarcate the boundaries of contact domains

A large fraction of peaks (38%) coincide with the corners of a contact domain — that is, the peak loci are located at domain boundaries (Figure 2.6A). Conversely, a large fraction of domains (39%) had peaks in their corner. Moreover, the appearance of a loop is usually (in 65% of cases) associated with the appearance of a domain demarcated by the loop. Because this configuration is so common, we use the term "loop domain" to refer to contact domains whose endpoints form a chromatin loop.

In some cases, adjacent loop domains (bounded by peak loci L_1-L_2 and L_2-L_3, respectively) exhibit transitivity — that is, L_1 and L_3 also correspond to a peak. This may indicate that the three loci simultaneously co-locate at a single spatial position. However, many peaks do not exhibit transitivity, suggesting that the corresponding loci do not co-locate. Figure 2.6B shows a region on chromosome 4 exhibiting both configurations.

We also found that overlapping loops are strongly disfavored: pairs of loops L_1-L_3 and L_2-L_4 (where L_1, L_2, L_3 and L_4 occur consecutively in the genome) are found 4-fold less often than expected under a random model.

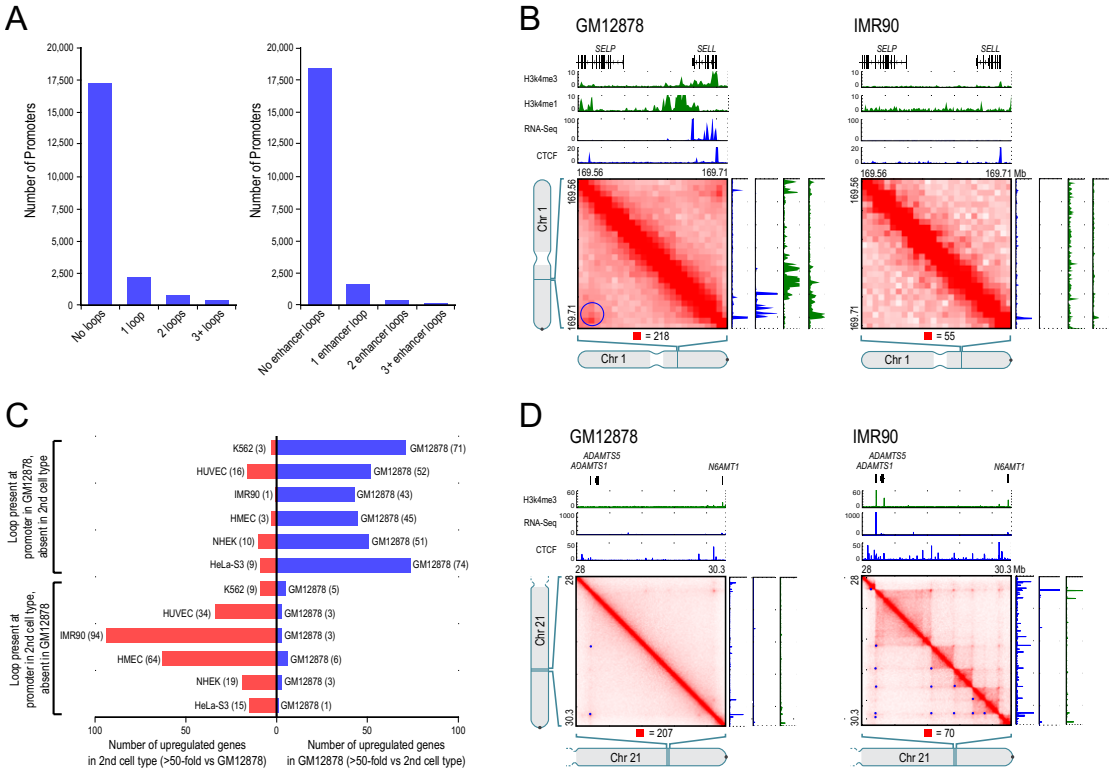


Figure 2.5: Loops between Promoters and Enhancers Are Strongly Associated with Gene Activation. (A) Histogram showing loop count at promoters (left); restricted to loops where the distal peak locus contains an enhancer (right). (B) Left: a loop in GM12878, with one anchor at the SELL promoter and the other at a distal enhancer. The gene is on. Right: the loop is absent in IMR90, where the gene is off. (C) Genes whose promoters participate in a loop in GM12878 but not in a second cell type are frequently upregulated in GM12878 and vice-versa. (D) Left: two loops in GM12878 are anchored at the promoter of the inactive ADAMTS1 gene. Right: a series of loops and domains appear, along with transitive looping. ADAMTS1 is on.

2.2.10 The vast majority of loops are associated with pairs of CTCF motifs in a convergent orientation

We next wondered whether peaks are associated with specific proteins. We examined the results of 86 ChIP-Seq experiments performed by ENCODE in GM12878. We found that the vast majority of peak loci are bound by the insulator protein CTCF (86%) and the cohesin subunits RAD21 (86%) and SMC3 (87%) (Figure 2.6C). This is consistent with numerous reports, using a variety of experimental modalities, that suggest a role for CTCF and cohesin in mediating DNA loops[64,115,136]. Because many of our loops demarcate domains, this observation is also consistent with studies suggesting that CTCF delimits structural and regulatory domains[36,41,148].

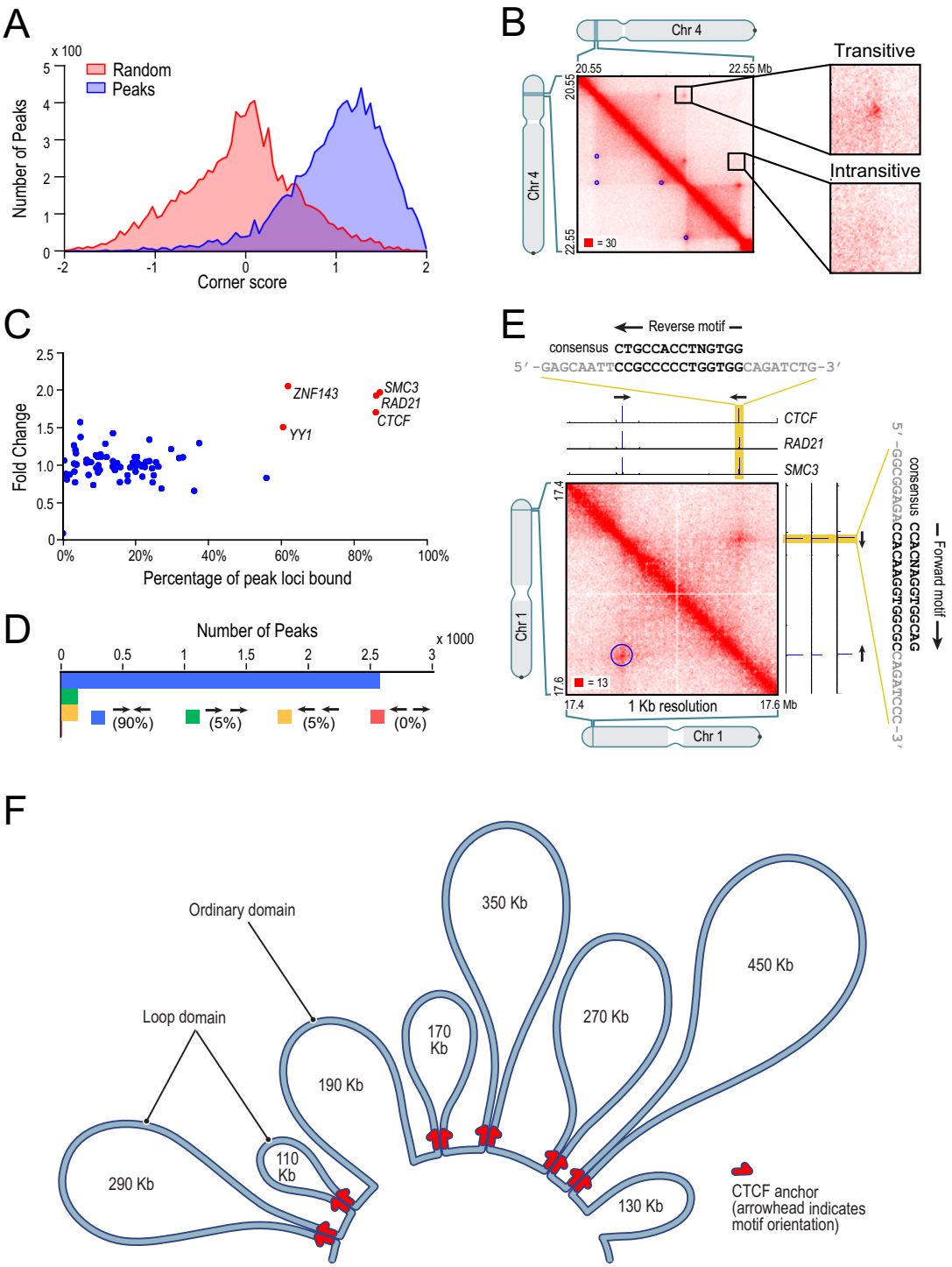
We found that most peak loci encompass a unique DNA site containing a CTCF-binding motif, to which all three proteins (CTCF, SMC3, and RAD21) were bound (5-fold enrichment). We were thus able to associate most of the peak loci (6991 of 12,903, or 54%) with a specific CTCF-motif "anchor".

The consensus DNA sequence for CTCF-binding sites is typically written as 5'-CCACNAGGTGGCAG-3'. Because the sequence is not palindromic, each CTCF motif has an orientation; we designate the consensus motif above as the "forward" orientation. Thus, a pair of CTCF sites on the same chromosome can have four possible orientations: (1) same direction on one strand; (2) same direction on the other strand; (3) convergent on opposite strands; and (4) divergent on opposite strands.

If CTCF sites were randomly oriented, one would expect all 4 orientations to occur equally often. But when we examined the 2857 peaks in GM12878 where the two corresponding peak loci each contained a single CTCF-binding motif, we found that the vast majority (90%) of motif pairs are convergent (Figure 2.6D,E). Overall, the presence,

Figure 2.6 *(following page)*: Many Loops Demarcate Contact Domains; the Vast Majority of Loops Are Anchored at a Pair of Convergent CTCF/RAD21/SMC3 Binding Sites. (A) Histograms of corner score for peak pixels versus random pixels with an identical distance distribution. (B) Contact matrix for chr4:20.55 Mb-22.55 Mb in GM12878, showing examples of transitive and intransitive looping behavior. (C) Percent of peak loci bound versus fold enrichment for 76 DNA-binding proteins. (D) The pairs of CTCF motifs that anchor a loop are nearly all found in the convergent orientation. (E) A peak on chromosome 1 and corresponding ChIP-seq tracks. Both peak loci contain a single site bound by CTCF, RAD21, and SMC3. The CTCF motifs at the anchors exhibit a convergent orientation. (F) A schematic rendering of a 2.1 Mb region on chromosome 20 (48.78-50.88 Mb). Eight domains tile the region, ranging in size from 110 kb to 450 kb; 95% of the region is contained inside a domain (contour lengths are shown to scale). Six of the eight domains are demarcated by loops between convergent CTCF- binding sites located at the domain boundaries. The other two domains are not demarcated by loops. The motifloop orientation is indicated by the direction of the arrow. Note that not every CTCF- binding site is shown.

Figure 2.6: (continued)



at pairs of peak loci, of bound CTCF sites in the convergent orientation was enriched 102-fold over random expectation. The convergent orientation was overwhelmingly more frequent than the divergent orientation, despite the fact that divergent motifs also lie on opposing strands: in GM12878, the counts were 2574-10 (257-fold enrichment, convergent vs. divergent); in IMR90, 1456-5 (291-fold); in HMEC, 968-11 (88-fold); in K562, 723-2 (362-fold); in HUVEC, 671-4 (168-fold); in HeLa, 301-3 (100-fold); in NHEK, 556-9 (62-fold); and in CH12, 625-8 (78-fold). This pattern suggests that a pair of CTCF sites in the convergent orientation is required for the formation of a loop.

The observation that looped CTCF sites occur in the convergent orientation also allows us to analyze peak loci containing multiple CTCF-bound motifs to predict which motif instance plays a role in a given loop. In this way, we can associate nearly two-thirds of peak loci (8,325 of 12,903, or 64.5%) with a single CTCF-binding motif.

The specific orientation of CTCF sites at observed peaks provides evidence that our peak calls are biologically correct. Because randomly chosen CTCF pairs would exhibit each of the four orientations with equal probability, the near-perfect association between our loop calls and the convergent orientation could not occur by chance ($p < 10\text{-}1900$, binomial distribution).

In addition, the presence of CTCF and RAD21 sites at many of our peaks provides an opportunity to compare our results to three recent ChIA-PET experiments reported by the ENCODE consortium (in GM12878 and K562) in which ligation junctions bound to CTCF (resp. RAD21) were isolated and analyzed. We found strong concordance with our results in all three cases[60,81].

2.2.11 The CTCF-binding exapted SINEB2 repeat in mouse shows preferential orientation with respect to loops

In mouse, we found that 7% of peak anchors lie within SINEB2 repeat elements containing a CTCF motif, which has been exapted to be functional. (The spread of CTCF binding via retrotransposition of this element, which contains a CTCF motif in its consensus sequence, has been documented in prior studies[20,128].) The CTCF motifs at peak anchors in SINEB2 elements show the same strong bias toward convergent orientation seen throughout the genome (89% are oriented towards the opposing loop anchor, vs. 94% genome-wide). The orientation of these CTCF motifs is aligned with the orientation of the SINEB2 consensus sequence in 97% of cases. This suggests that exaptation of a CTCF in a SINEB2 element is more likely when the orientation of the inserted SINEB2 is compatible with local loop structure.

2.2.12 Diploid Hi-C maps reveal homolog-specific features, including imprinting-specific loops and massive domains and loops on the inactive X chromosome

Because many of our reads overlap SNPs, it is possible to use GM12878 phasing data[52,91] to assign contacts to specific chromosomal homologs (Figure 2.7A). Using these assignments, we constructed a "diploid" Hi-C map of GM12878 comprising both maternal (238M contacts) and paternal (240M) maps.

For autosomes, the maternal and paternal homologs exhibit very similar inter- and intra-chromosomal contact profiles (Pearson's R>.998). One interchromosomal difference was notable: an elevated contact frequency between the paternal homologs of chromosome 6 and 11 that is consistent with an unbalanced translocation fusing chr11q:73.5

Mb and all distal loci (a stretch of over 60 Mb) to the telomere of chromosome 6p (Figure 2.7B). The signal intensity suggests that the translocation is present in between 1.2% and 5.6% of our cells. We tested this prediction by karyotyping 100 GM12878 cells using Giemsa staining and found 3 abnormal chromosomes, each showing the predicted translocation, der(6)t(6,11)(pter;q). The Hi-C data reveal that the translocation involves the paternal homologs, which cannot be determined with ordinary cytogenetic methods.

We also observed differences in loop structure between homologous autosomes at some imprinted loci. For instance, the H19/Igf2 locus on chromosome 11 is a well-characterized case of genomic imprinting. In our unphased maps, we clearly see two loops from a single distal locus at 1.72 Mb (which binds CTCF in the forward orientation) to loci located near the promoters of both H19 and Igf2 (both of which bind CTCF in the reverse orientation, i.e., the above consensus motif lies on the opposite strand; see Fig. 7C). We refer to this distal locus as the H19/Igf2 Distal Anchor Domain (HIDAD). Our diploid maps reveal that the loop to the H19 region is present on the maternal chromosome (from which H19 is expressed), but the loop to the Igf2 region is absent or greatly attenuated. The opposite pattern is found on the paternal chromosome (from which Igf2 is expressed).

Pronounced differences were seen on the diploid intra-chromosomal maps of chromosome X. The paternal X chromosome, which is usually inactive in GM12878, is partitioned into two massive domains (0-115 Mb and 115-155.3 Mb). These "superdomains" are not seen in the active, maternal X (Fig. 7D). When we examined the unphased maps of chromosome X for the karyotypically normal female cell lines in our study (GM12878, IMR90, HMEC, NHEK), the superdomains on X were evident, although the signal was attenuated due to the superposition of signals from active and inactive X chromosomes.

When we examined the male HUVEC cell line and the haploid KBM7 cell line, we saw no evidence of superdomains.

Interestingly, the boundary between the superdomains (ChrX: 115 Mb +/- 500 kb) lies near the macrosatellite repeat DXZ4 (ChrX: 114,867,433—114,919,088) near the middle of Xq. DXZ4 is a CpG-rich tandem repeat that is conserved across primates and monkeys and encodes a long non-coding RNA. In males and on the active X, DXZ4 is heterochromatic, hyper-methylated and does not bind CTCF. On the inactive X, DXZ4 is euchromatic, hypo-methylated, and binds CTCF. DXZ4 has been hypothesized to play a role in reorganizing chromatin during X inactivation[24].

There were also significant differences in loop structure between the chromosome X homologs. We observed 27 large "superloops," each spanning between 7 and 74 Mb, present only on the inactive X chromosome in the diploid map (Fig. 7E). The superloops were also seen in all 4 unphased maps from karyotypically normal XX cells, but were absent in unphased maps from X0 and XY cells. Two of the superloops (chrX:56.8 Mb-DXZ4 and DXZ4-130.9 Mb) were reported previously, in a locus-specific study[63].

Like the peak loci of most other loops, nearly all the superloop anchors bind CTCF (23 of 24). The six anchor regions most frequently associated with superloops are large (up to 200 kb). Four of these anchor regions contain whole lncRNA genes: loc550643; XIST; DXZ4; and FIRRE. Three (loc550643, DXZ4, and FIRRE) contain CTCF-binding tandem repeats that only bind CTCF on the inactive homolog.

2.3 Discussion

Using the in situ Hi-C protocol, we probed genomic architecture with high resolution; in the case of GM12878 lymphoblastoid cells, better than 1 kb. We observe the presence of contact domains that were too small (median length = 185 kb) to be seen in previous maps. Loci within a domain interact frequently with one another, have similar patterns of chromatin modifications, and exhibit similar long-range contact patterns. Domains tend to be conserved across cell types and between human and mouse. When the pattern of chromatin modifications associated with a domain changes, the domain's long-range contact pattern also changes. Domains exhibit at least six distinct patterns of long-range contacts (subcompartments), which subdivide the two compartments that we previously reported based on low resolution data. The subcompartments are each associated with distinct chromatin patterns. It is possible that the chromatin patterns play a role in bringing about the long-range contact patterns, or vice-versa.

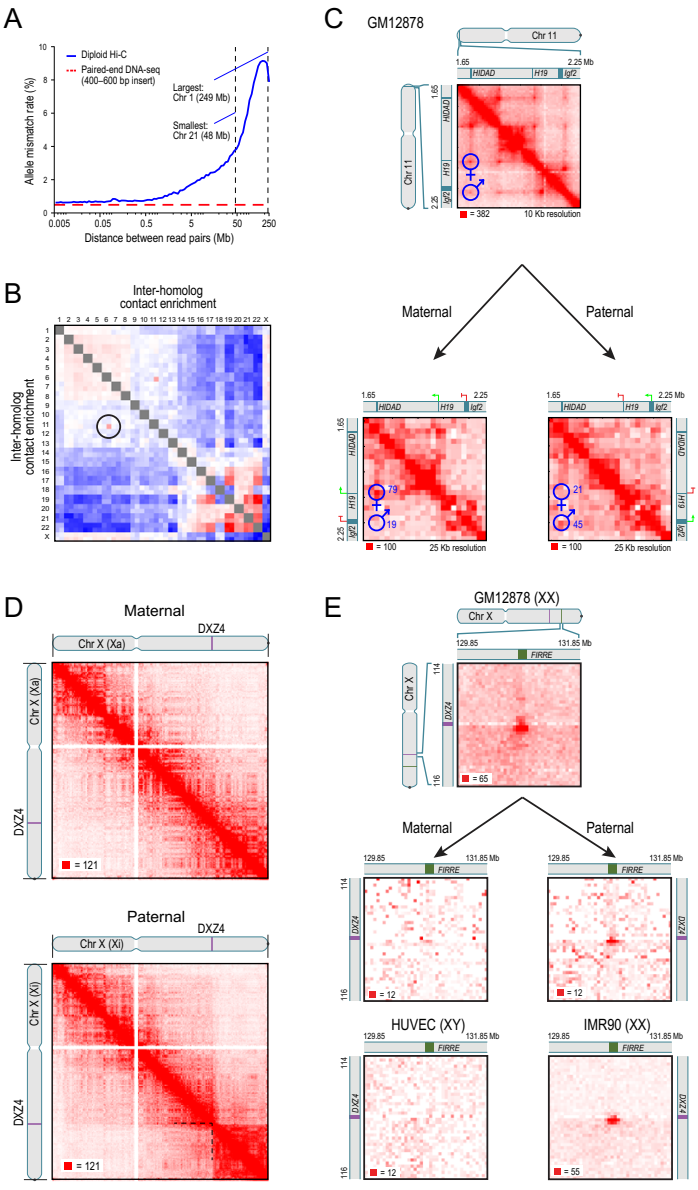
Our data also make it possible to create a genome-wide catalog of chromatin loops. We identified loops by looking for pairs of loci that have significantly more contacts with one another than they do with other nearby loci. In our densest map (GM12878), we observe 9448 loops.

The loops reported here have many interesting properties. Most loops are short (<2 Mb) and strongly conserved across cell types and between human and mouse. Promoter-enhancer loops are common and associated with gene activation. Loops tend not to overlap; they often demarcate contact domains, and may establish them. CTCF and the cohesin subunits RAD21 and SMC3 associate with loops; each of these proteins is found at over 86% of loop anchors.

The most striking property of loops is that the pair of CTCF motifs present at the

Figure 2.7 *(following page)*: Diploid Hi-C Maps Reveal Superdomains and Superloops Anchored at CTCF-Binding Tandem Repeats on the Inactive X Chromosome. (A) The frequency of mismatch (maternal-paternal) in SNP allele assignment versus distance between two paired read alignments. Intrachromosomal read pairs are overwhelmingly intramolecular. (B) Preferential interactions between homologs. Left/top is maternal; right/bottom is paternal. The aberrant contact frequency between 6/ paternal and 11/ paternal (circle) reveals a translocation. (C) Top: in our unphased Hi-C map of GM12878, we observe two loops joining both the promoter of the maternally-expressed H19 and the promoter of the paternally-expressed Igf2 to a distal locus, HIDAD. Using diploid Hi-C maps, we phase these loops: the HIDAD-H19 loop is present only on the maternal homolog (left) and the HIDAD-Igf2 loop is present only on the paternal homolog (right). (D) The inactive (paternal) copy of chromosome X (bottom) is partitioned into two massive "superdomains" not seen in the active (maternal) copy (top). DXZ4 lies at the boundary. (E) The "superloop" between FIRRE and DXZ4 is present in the unphased GM12878 map (top), in the paternal GM12878 map (middle right), and in the map of the female cell line IMR90 (bottom right); it is absent from the maternal GM12878 map (middle left) and the map of the male HUVEC cell line (bottom left).

Figure 2.7: (continued)



loop anchors occurs in a convergent orientation in >90% of cases (vs. 25% expected by chance). The importance of motif orientation between loci that are separated by, on average, 360 kb is surprising and must bear on the mechanism by which CTCF and cohesin form loops, which seems likely to involve CTCF dimerization. Experiments in which the presence or orientation of CTCF sites is altered may enable the engineering of loops, domains, and other chromatin structures.

It is interesting to compare our results to those seen in previous reports. The contact domains we observe are similar in size to the "physical domains" that have been reported in Hi-C maps of Drosophila[134] and to the "topologically constrained domains" (mean length: 220 kb) whose existence was demonstrated in the 1970s and 1980s in structural studies of human chromatin[31,143,153]. On the other hand, the domains we observe are much smaller than the TADs (1 Mb)[41] that have been reported in humans and mice on the basis of lower-resolution contact maps. This is because detecting TADs involves detection of domain boundaries. With higher resolution data, it is possible to detect additional boundaries beyond those seen in previous maps. Interestingly, nearly all the boundaries we observe are associated with either a subcompartment transition (which occur roughly every 300 kb), or a loop (which occur roughly every 200 kb); and many are associated with both.

Our annotation identifies many fewer loops than were reported in several recent high-throughput studies, despite the fact that we have more data. The key reason is that we call peaks only when a pair of loci shows elevated contact frequency relative to the local background — that is, when the peak pixel is enriched as compared to other pixels in its neighborhood. In contrast, prior studies have defined peaks by comparing the contact frequency at a pixel to the genome-wide average[69,124]. This latter definition is problematic because many pixels within a domain can be annotated as peaks despite showing

no local increase in contact frequency. Papers using the latter definition imply the existence of more than 100,000 loops (1187 loops were reported in 1% of the genome[124]) or even more than 1 million loops (reported in a genome-wide Hi-C study[69]). The vast majority of the loops annotated by these papers show no enrichment relative to the local background when examined one-by-one, and no enrichment with respect to any published Hi-C dataset when analyzed using APA. This suggests that these peak annotations may correspond to pairs of loci that lie in the same domain or compartment, but rarely correspond to loops.

We created diploid Hi-C maps by using polymorphisms to assign contacts to distinct chromosomal homologs. We found that the inactive X chromosome is partitioned into two large superdomains whose boundary lies near the locus of the lncRNA DXZ4. We also detect a network of long-range superloops, the strongest of which are anchored at locations containing lncRNA genes (loc550643, XIST, DXZ4, and FIRRE). With the exception of XIST, all of these lncRNAs contain CTCF-binding tandem repeats that bind CTCF only on the inactive X.

In our original report on Hi-C, we observed that Hi-C maps can be used to study physical models of genome folding, and we proposed a fractal globule model for genome folding at the megabase scale. The kilobase-scale maps reported here allow the physical properties of genome folding to be probed at much higher resolution. We will report such studies elsewhere.

Just as loops bring distant DNA loci into close spatial proximity, we find that they bring disparate aspects of DNA biology — domains, compartments, chromatin marks, and genetic regulation — into close conceptual proximity. As our understanding of the physical connections between DNA loci continues to improve, our understanding of the relationships between these broader phenomena will deepen.

2.4 Experimental Procedures

2.4.1 In situ Hi-C Protocol

All cell lines were cultured following the manufacturer's recommendations. Two to five million cells were crosslinked with 1% formaldehyde for 10 minutes at room temperature. Nuclei were permeabilized. DNA was digested with 100 units of MboI, and the ends of restriction fragments were labeled using biotinylated nucleotides and ligated in a small volume. After reversal of crosslinks, ligated DNA was purified and sheared to a length of roughly 400 bp, at which point ligation junctions were pulled down with streptavidin beads and prepped for Illumina sequencing. Dilution Hi-C was performed as in Erez Lieberman-Aiden, et al.[83].

2.4.2 3D-FISH

3D DNA FISH was performed as in Brian Beliveau, et al.[12], with minor modifications.

2.4.3 Hi-C Data Pipeline

All sequence data was produced using Illumina paired-end sequencing. We processed data using a custom pipeline that was optimized for parallel computation on a cluster. The pipeline uses BWA[82] to map each read end separately to the b37 or mm9 reference genomes; removes duplicate and near-duplicate reads; removes reads that map to the same fragment; and filters the remaining reads based on mapping quality score. Contact matrices were generated at base pair delimited resolutions of 2.5 Mb, 1 Mb, 500 kb, 250 kb, 100 kb, 50 kb, 25 kb, 10 kb, and 5 kb, as well as fragment-delimited resolutions of 500f, 200f, 100f, 50f, 20f, 5f, 2f, and 1f. For our largest maps, we also generated a 1 kb contact matrix. Normalized contact matrices are produced at all resolutions using P.

Knight and D. Ruiz[76].

2.4.4 Annotation of Domains: Arrowhead

To annotate domains, we apply a novel "arrowhead" transformation, defined as $A_{i,i+d} = (M^*_{i,i-d} - M^*_{i,i+d})/(M^*_{i,i-d} + M^*_{i,i+d})$. M^* denotes the normalized contact matrix. This is equivalent to calculating a matrix equal to -1*(observed/expected-1), where the expected model controls for local background and distance from the diagonal in the simplest possible way: the "expected" value at i,i+d is simply the mean of the observed values at i,i-d and i,i+d. $A_{i,i+d}$ will be strongly positive if locus i-d is inside a domain and locus i+d is not. If the reverse is true, $A_{i,i+d}$ will be strongly negative. If the loci are both inside or both outside a domain, $A_{i,i+d}$ will be close to zero. Consequently, if there is a domain at [a,b], we find that A takes on very negative values inside a triangle whose vertices lie at [a,a], [a,b], and [(a+b)/2,b], and very positive values inside a triangle whose vertices lie at [(a+b)/2,b], [b,b], and [b,2b-a]. The size and positioning of these triangles creates the arrowhead-shaped feature that replaces each domain in M^*. A "corner score" matrix, indicating each pixel's likelihood of lying at the corner of a domain, is efficiently calculated from the arrowhead matrix using dynamic programming.

2.4.5 Assigning loci to subcompartments

To cluster loci based on long-range contact patterns, we constructed a 100 kb resolution interchromosomal contact matrix such that loci from odd chromosomes appeared on the rows, and loci from even chromosomes appeared on the columns. (Intrachromosomal data, and data involving chromosome X, were excluded.) We cluster this matrix using the Python package scikit. For subcompartment B4, the 100 kb interchromosomal matrix for chromosome 19 was constructed and clustered separately, using the same

procedure.

2.4.6 Annotation of Peaks: HiCCUPS

Our peak-calling algorithm examines each pixel in a Hi-C contact matrix and compares the number of contacts in the pixel to the number of contacts in a series of regions surrounding the pixel. The algorithm thus identifies "enriched pixels" $M^*_{i,j}$ where the contact frequency is higher than expected, and where this enrichment is not the result of a larger structural feature. For instance, we rule out the possibility that the enrichment of pixel $M^*_{i,j}$ is the result of L_i and L_j lying in the same domain by comparing the pixel's contact count to an expected model derived by examining the "lower-left" neighborhood. (The "lower-left" neighborhood samples pixels $M^*_{i',j'}$ where $i \leq i' \leq j' \leq j$; if pixel $M^*_{i,j}$ is in a domain, these pixels will necessarily be in the same domain.) We require that the pixel being tested contain at least 50% more contacts than expected based on the lower-left neighborhood, and that the enrichment be statistically significant after correcting for multiple hypothesis testing (FDR<10%). The same criteria are also applied to three other neighborhoods. Thus, to be labeled an enriched pixel, a pixel must be significantly enriched relative to four neighborhoods: (i) pixels to its lower-left; (ii) pixels to its left and right; (iii) pixels above and below; and (iv) a donut surrounding the pixel of interest (Figure 2.4A). The resulting enriched pixels tend to form contiguous interaction regions comprising 5-20 pixels each. We define the "peak pixel" (or simply the "peak") to be the pixel in an interaction region with the most contacts.

Because of the enormous number of pixels that must be examined, this calculation requires weeks of CPU time to execute. (For instance, at a matrix resolution of 5 kb, the algorithm must be run on over 40 billion pixels.) To accelerate it, we created a highly parallelized implementation using general-purpose graphical processing units, resulting

in a 200-fold speedup.

2.4.7 Aggregate Peak Analysis

We perform APA on 10 kb resolution contact matrices. To measure the aggregate enrichment of a set of putative peaks in a contact matrix, we plot the sum of a series of submatrices derived from that contact matrix. Each of these submatrices is a 210 kb x 210 kb square centered at a single putative peak in the upper triangle of the contact matrix. The resulting APA plot displays the total number of contacts that lie within the entire putative peak set at the center of the matrix; the entry immediately to the right of center corresponds to the total number of contacts in the pixel set obtained by shifting the peak set 10 kb to the right; the entry two positions above center corresponds to an upward shift of 20 kb, and so on. Focal enrichment across the peak set in aggregate manifests as larger values at the center of the APA plot. The APA plots shown only include peaks whose loci are at least 300 kb apart.

3

Re-analysis of Prior Loop Lists

In our work, we annotated loops in each of the cell types we studied. We then sought to validate these loop lists by examining their enrichment in other Hi-C experiments performed in the same cell type. For the GM12878 cell line, we created two deep contact maps, and so we could annotate peaks in both experiments and then compare the reproducibility. However, we also wanted to examine the enrichment of these loops on previously published Hi-C maps, which had much less data and were therefore much sparser than our in situ maps. Individual loops are frequently impossible to discern in such sparse maps, so comparison of two annotated lists was not an option.

To get around this, we developed Aggregate Peak Analysis (APA) expressly for this purpose. Explained in more detail in the following chapter, APA tests for the aggregate

enrichment of an entire set of putative two-dimensional peaks. APA is carried out by examining all of the putative loops of a loop list in a relevant Hi-C contact map and calculating their collective enrichment. The putative loops may have been called in a different Hi-C map in the same cell type, or may have been imputed via an orthogonal technique. The APA score is computed by calculating the fold enrichment of all of the loops relative to their local neighborhoods. If the putative loop list contains mostly true loops, the enrichment will be large, resulting in a high APA score (significantly above 1). However, if the putative loop lists contains mostly false positives, there will be no aggregate enrichment, and the APA score will be close to 1.

We used APA to test for the enrichment both of the loop lists that we ourselves annotated using in situ data, and also of previously published loop lists in the literature. Notably, we did not find an instance in which one of the in situ Hi-C peak lists that we annotate failed to be validated, in aggregate, when examined with respect to a successful Hi-C experiment in the same cell type. In contrast, we found that previously annotated loops list in the literature from a number of other experiments, which posited close to a million loops in the human genome, are dominated by false positives and not validated via APA.

3.1 APA of in situ Hi-C Peak Lists

Using APA, we found that our peak annotations exhibit focal enrichment in all published Hi-C datasets. APA was performed on all published lymphoblast Hi-C datasets using our GM12878 peak annotations. The resulting APA plots are shown in Fig. Fig. 3.1E. Focal enrichment was seen in all cases, even on sparse maps with <10M reads. We also performed APA on maps we constructed using the dilution Hi-C protocol and maps we constructed using the in situ Hi-C protocol without crosslinking (Fig. Fig. 3.1E).

Notably, even without the use of crosslinking, we observed a robust enrichment (HIC043, APA score: 2.838, Z-score: 26.241; HIC044, APA score: 2.030, Z-score: 14.586; HIC045, APA score: 3.063, Z-score: 8.814; HIC046, APA score: 2.105, Z-score: 15.962; HIC047, APA score: 2.114, Z-score: 8.060). As such, our loops cannot be a result of crosslinking biases. Finally, we performed APA on 107 additional coarse resolution experiments covering a variety of experimental conditions; enrichment was seen in every case.

Additionally, APA was performed on the relevant published Hi-C maps or our own dilution maps for our other in situ peak lists: IMR90 peaks (APA performed on 3 maps), K562 peaks (1 map), HMEC peaks (1 map), HUVEC peaks (1 map), NHEK peaks (1 map), and mouse lymphoblastoid peaks (2 maps). Again, focal enrichment was seen in all cases (see Fig. 3.1F).

3.2 APA of Previous Loop Annotation Lists

Several recent large-scale studies have endeavored to call loops in a high-throughput fashion, reporting far more loops than we identify in our in situ Hi-C maps[69,81,124]. For example, an ENCODE study by Sanyal et al. in GM12878 used 5C to call 1187 putative peaks in 1% of the human genome[124], suggesting the existence of 120,000 peaks genome-wide. A CHIA-PET experiment in K562 cells by Li et al., also reported by ENCODE, observed 126,886 putative peaks anchored at PolII sites alone[81]. A dilution Hi-C study of IMR90 by Jin et al. called 1.1 million putative peaks across the genome[69]. In contrast, although our GM12878 map contains 25-fold more contact data than the Jin et al. paper, we find only about 10,000 peaks genome-wide (Fig. 3.2A,B).

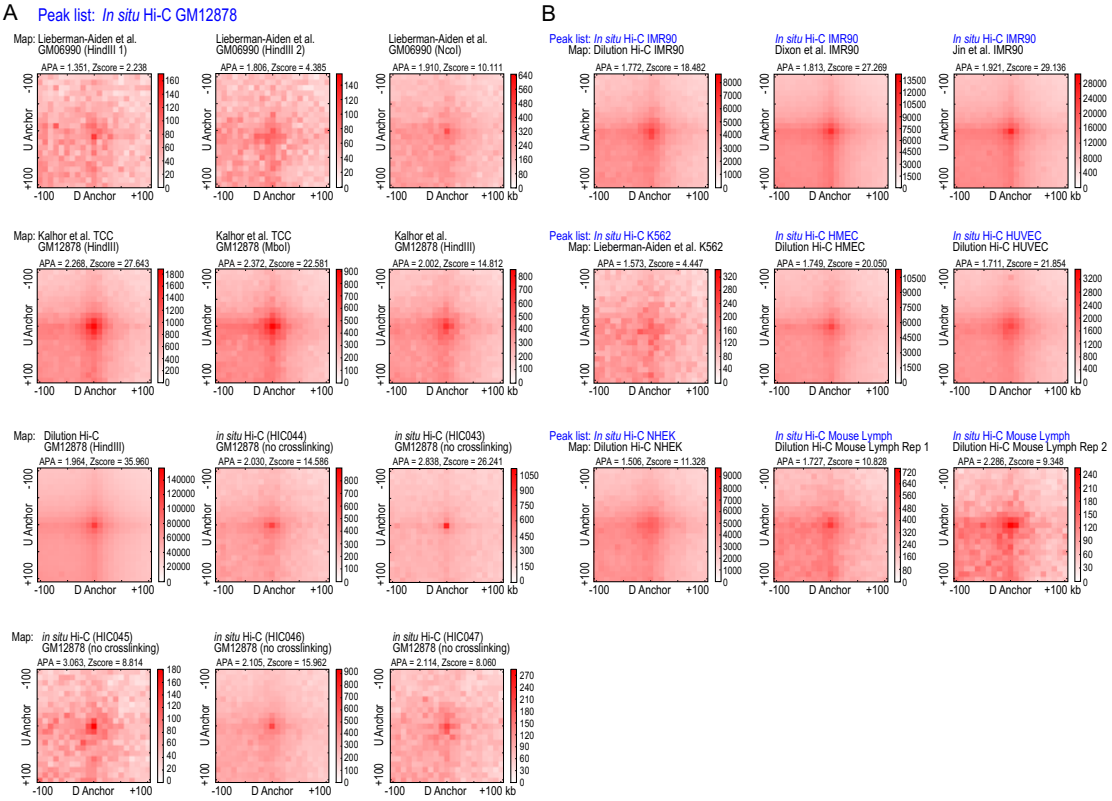


Figure 3.1: APA validates in situ loop lists. (A) We show APA plots for our in situ GM12878 peak list on every published human lymphoblastoid Hi-C contact map as well as additional lymphoblastoid maps generated in this study. In all cases the aggregate focal enrichment of our peaks can be seen. (B) We show APA plots for our in situ IMR90, K562, HMEC, HUVEC, NHEK, and mouse lymphoblastoid peak lists against all available Hi-C maps in each cell line. In all cases focal enrichment is observed.

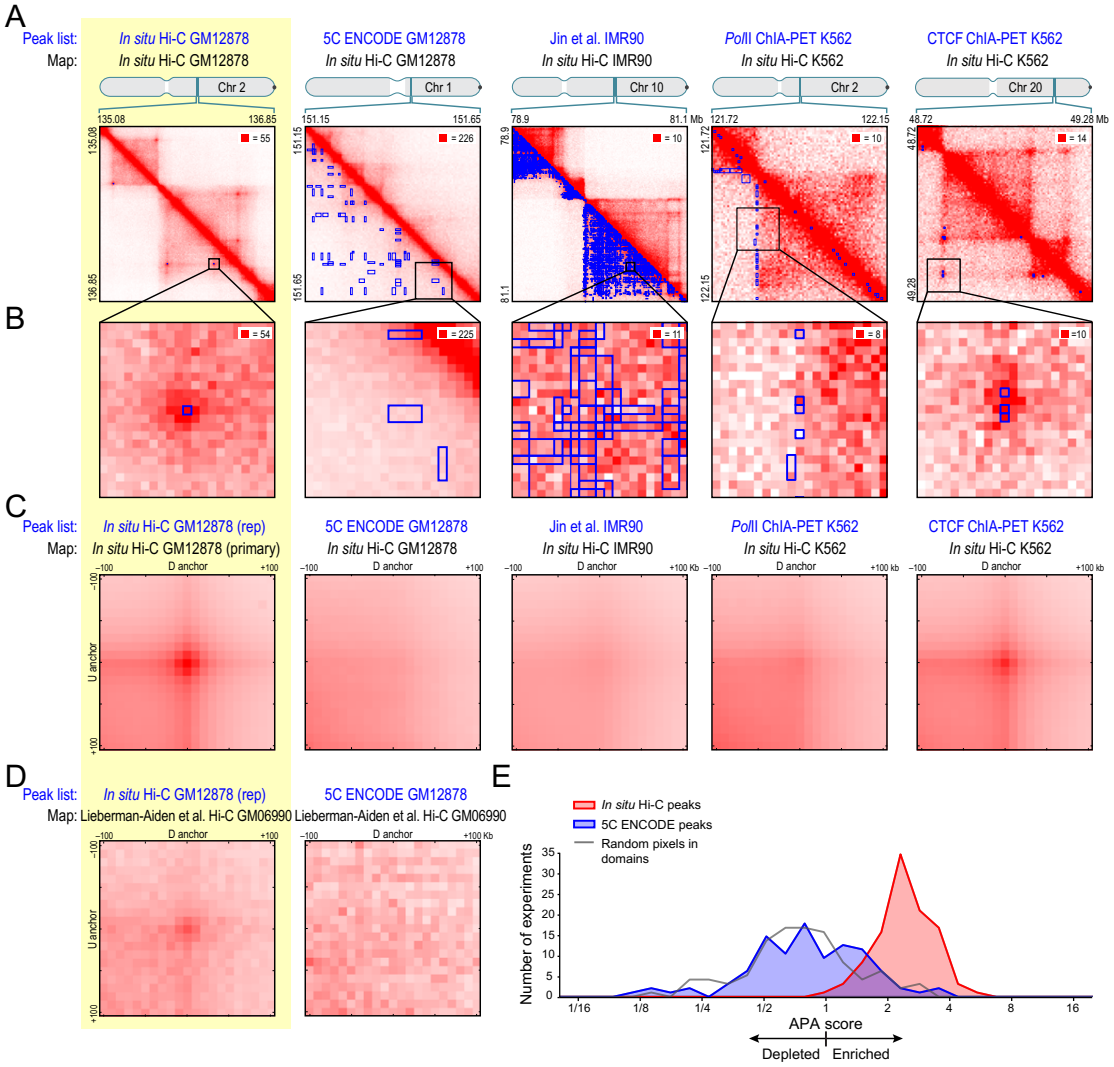
We suspected that this disparity might be due to the methodology used for calling peaks in these earlier studies. As noted above, we were careful to call peaks only when a pair of loci showed elevated contact frequency relative to other nearby loci; i.e., we only call a pixel a peak if it is enriched above the other pixels in its neighborhood. In contrast, the studies noted above call a peak by comparing contact frequency to a genome-wide average. Using such a procedure, it is difficult to tell whether the high contact frequency observed at a pixel represents a specific interaction between two loci (i.e., a loop), or merely an overall enrichment in contact frequency across all loci in a local region (for instance, a domain). In fact, as a control, when we called loops in in situ Hi-C data by comparing each pixel to the genome-wide average, over 2 million putative loops were detected, nearly all of which were not enriched above the local neighborhood and thus are not biologically meaningful.

We therefore sought to assess the accuracy of earlier loop annotations. First, we re-examined the putative peaks annotated by Sanyal et al. in GM12878 using our maps, which had deep coverage of the corresponding pixels (median # of contacts: 99.9). When we visually inspected every pixel in a 1.7Mb region containing 103 putative peaks (Fig. 3.2A,B), we found only 2 peaks that showed enrichment over local background. When we computationally analyzed 1114 putative peaks in this list (excluding peaks between loci closer than 35Kb), we found only 52 peaks that were enriched above local background. When we compared the putative peak set using APA to our three high-coverage GM12878 maps (primary, replicate, and dilution) we did not observe statistically significant enrichment in a single case (Fig. 3.2C-D).

We performed 100 additional experiments in GM12878 using the dilution Hi-C, in situ Hi-C, and TCC protocols, varying a wide variety of experimental parameters, and using many separate biological samples. We produced Hi-C maps for each of these ex-

Figure 3.2 *(following page)*: Previous high-throughput loop annotations are dominated by false positives. Comparison of loop lists reported in this paper (1st column) with earlier studies based on 5C in GM12878 (2nd column), Hi-C in IMR90 (3rd column), CHIA-PET in K562 using PolII as bait (4th column) and using CTCF as bait (5th column). In (A), each putative loop list (blue) is superimposed on the corresponding in situ Hi-C map at 5Kb resolution. In (B), we zoom in on a subregion. The loops reported in this study and using CTCF/CHIA-PET show clear enrichment over local background. Putative loops annotated in other previous studies do not, although they can correspond to larger features such as the edges or interiors of domains. This implies that these putative loops are false positives. (C) APA analysis of various loop lists. If the fraction of false positives is small, there is strong focal enrichment at the center of the aggregate heatmap. Loop calls generated using our replicate GM12878 map show strong focal enrichment in the primary GM12878 map. Similar results are seen for CTCF/CHIA-PET loop calls. Focal enrichment is not seen for previously reported lists based on 5C, PolII/CHIA-PET, and dilution Hi-C. Note that false positives along domain edges lead to strong enrichment in the lower-left of an APA plot, but not in the center. (D) Left: APA analysis of loops from our replicate GM12878 using data from our original GM06990 dilution Hi-C maps (Lieberman-Aiden, et al.) containing only 36M contacts. Focal enrichment is apparent. Right: putative loops from ENCODE's GM2878 5C experiment show no focal enrichment in the original Hi-C experiment (APA score: 0.88). (E) Histogram showing APA analyses comparing our in situ Hi-C loop list (red) and ENCODE's 5C loop list (blue) with respect to 100 contact maps generated using many variants of the Hi-C protocol. The results for our loop list are tightly distributed around 2, suggesting 2-fold enrichment above local background in a typical map. The ENCODE/5C loops show no enrichment, closely resembling results for a "control loop list" (grey) generated by selecting pixels at random inside domains.

Figure 3.2: (continued)



periments, containing, on average, 1.2M contacts. When we examined our GM12878 in situ loop list using APA, we found enrichment in the center of the APA plot (APA score > 1) with respect to every one of the 100 experiments (Fig. 3.2E). The results for our loop list are tightly distributed around 2, suggesting approximately 2-fold enrichment above local background in a typical map. The list of putative loops generated by ENCODE in GM12878 (Fig. 3.2E, blue) is centered at unity, and did not generate a statistically significant enrichment above local background in a single case ($p<.05$ after Bonferroni correction). The histogram for ENCODE's 5C putative loops closely resembles a histogram created when we repeated this procedure using a "control loop list," (grey) which was generated by selecting pixels at random inside contact domains.

We also compared the ENCODE putative loop list to all six previously published Hi-C lymphoblastoid maps by various groups; similarly we did not observe statistically significant enrichment in a single case (Fig 3.3).

We repeated the above procedure on the peaks annotated by Li et al. using PolII/CHIA-PET in K562, and on the peaks annotated by Jin et al. using Hi-C in IMR90. The results were very similar (Fig. 3.2A-C). A recent CHIA-PET experiment in K562 that used CTCF as bait[81], however, did show focal enrichment over local background, both when we examined peaks individually and when we performed APA.

Taken together, these findings suggest that CTCF/CHIA-PET experiments provide useful data when annotating mammalian loops. However, they suggest that the putative loops reported by the other experiments shown, which were inferred based on comparison to a genome-wide background, may represent pairs of loci that lie in the same domain or compartment, but are not pairs of loci that show enhanced contacts relative to the local background. Accordingly, these locus pairs should not be regarded as the specific anchors for loops.

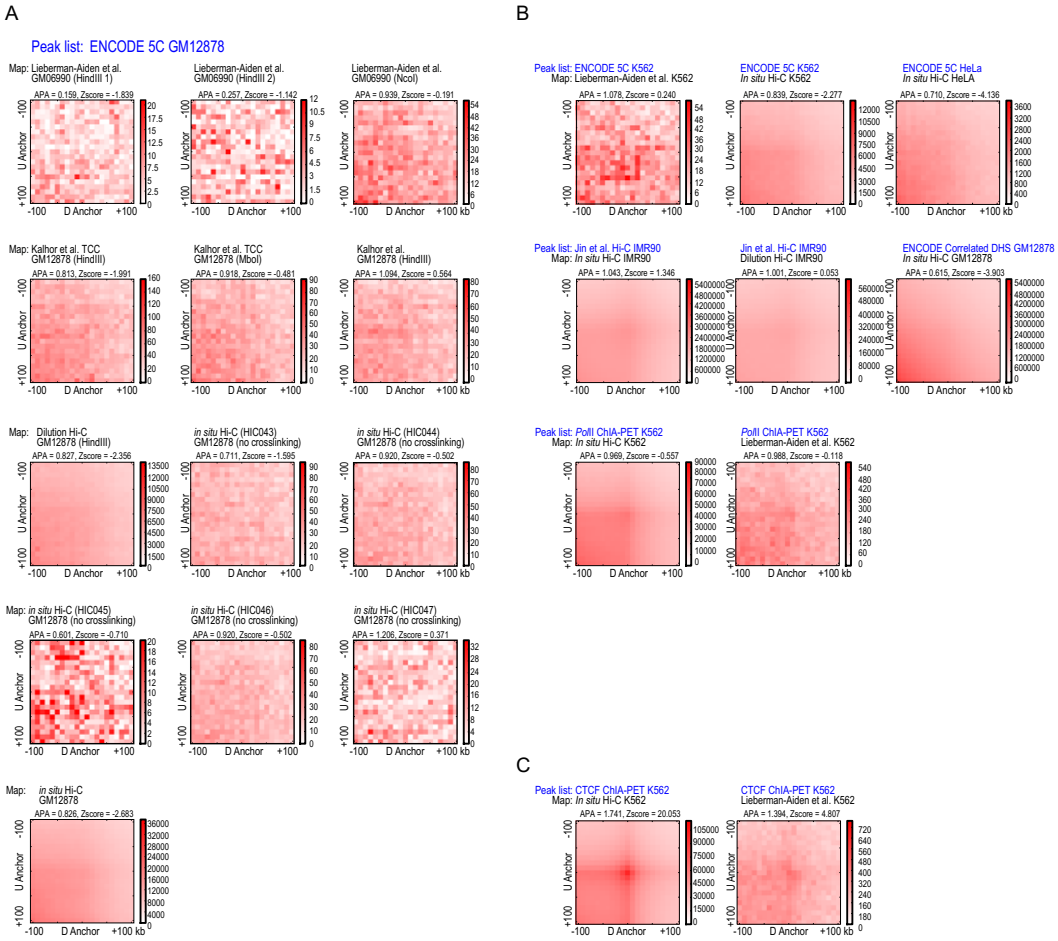


Figure 3.3: APA does not so enrichment for previous loop lists.(A) We show APA plots for the ENCODE 5C GM12878 peak list on every published human lymphoblastoid Hi-C contact map as additional lymphoblastoid maps generated in this study. No aggregate focal enrichment can be seen. (B) We show APA plots for the ENCODE 5C K562 peak list, the ENCODE 5C HeLa peak list, the Jin et al. Hi-C IMR90 peak list, the ENCODE correlated GM12878 DHS pairs list, and the PolII ChIA-PET K562 peak list against all available Hi-C maps in each cell line. No aggregate focal enrichment can be seen. (C) APA plots for the CTCF ChIA-PET K562 peak list in the two K562 Hi-C maps indicate that these peaks do exhibit focal enrichment.

4

New Quantitative Methods for Hi-C Analysis

This chapter details some of the novel methods we developed to analyze high-resolution contact maps. Here we discuss the Arrowhead algorithm, used to annotate contact domains, and the resulting properties of the contact domains we find; the long-range pattern clustering algorithm, and some of the characteristics of the resulting subcompartments we annotate; and Aggregate Peak Analysis, used to validate loop annotation lists via an aggregate measure of enrichment.

4.1 The Arrowhead Algorithm

4.1.1 Motivation and related work

The formation of square megadomains along the diagonal of a contact map is a striking feature that was apparent in our 2009 maps, and which we explained in terms of compartmentalization[83]. Subsequent work has highlighted the computational problem of identifying domains (which manifest as squares along the diagonal of a contact map) as distinct from the problem of identifying compartments[41,133].

The fact that domains manifest as squares along the diagonal of a contact map suggests that they should be straightforward to identify. In practice, however, the identification of domains is tricky. This is due to experimental factors such as noise and inadequate coverage. It is also because of the intrinsic difficulty of the problem: the decline in contact frequency at domain edges can be subtle, and the very rapid decline in contact probability observed as one moves away from the diagonal of a contact map is a major confound for most approaches.

Nevertheless, several methods exist for identifying domains. Notably, Dixon et al.[41] defined a directionality index (DI), which measures the tendency of a locus to interact with upstream vs. downstream sites. This is useful for identifying domains because the upstream boundary of a domain should prefer to interact with downstream loci, and vice-versa.

4.1.2 Description of Arrowhead transformation

The arrowhead transformation (see Figure 4.1A-F) is a matrix transformation defined as $A_{i,i+d} = (M^*_{i,i-d} - M^*_{i,i+d)}/(M^*_{i,i-d} + M^*_{i,i+d})$. This transformation can be thought of as equivalent to calculating a matrix equal to $-1 * ((observed/expected) - 1)$, where

the expected model controls for local background and distance from the diagonal in the simplest possible way: the "expected" value at $i, i+d$ is simply the mean observed value at $i, i-d$ and $i, i+d$. By choosing variants on this expected model, one can create a family of related transformations with similar properties. Alternatively, one can think of $A_{i,i+d}$ as a measurement of the directionality preference of locus i, restricted to contacts at a linear distance of d.

Consider the behavior of this transformation when a domain is present in M^* between locus a and locus b (i.e., there is a square of enriched contact frequency whose vertices lie at (a,a), (a,b), (b,b), and (b,a). $A_{i,i+d}$ will be strongly positive if and only if locus $i-d$ is inside the domain (i.e., in the range $[a,b]$) and locus $i+d$ is not. $A_{i,i+d}$ will be strongly negative when locus $i+d$ is inside the domain and locus $i-d$ is not. If both loci are inside the domain, or both loci are outside the domain, $A_{i,i+d}$ will be close to zero. (Note that this behavior also exploits the fact that one typically observes squares of depleted contact frequency adjacent to domains.)

Thus, the general behavior of the arrowhead matrix A can be seen by solving a series of simple inequalities that follow from the above statement. If we think of the solution geometrically, we see that A takes on very negative values inside an "upper" triangle $U_{a,b}$, whose vertices lie at (a,a), (a,b), and $((a+b)/2, b)$. We also see that A takes on very positive values inside a "lower" triangle $L_{a,b}$, whose vertices lie at $((a+b)/2, b)$, (b,b), and $(b, 2b-a)$. Everywhere else, the entries of A are close to zero.

One can think of the "upper" and "lower" triangles as a smear that exaggerates the original edges of the domain, making these features easier to detect. The negligible values seen everywhere else also have an important consequence: they replace the steep decline seen inside a domain in M^* - which tends to confound feature detection algorithms - with a relatively constant region in A. Because of the mirror symmetry of the

matrix A, the effect of the transformation, when examined as a whole, is to transform an (abnormally-hard-to-annotate) square feature into a (relatively-easy-to-annotate) arrowhead shaped feature. See Figure 4.1A-F.

4.1.3 Arrowhead scoring

The goal of our algorithm is to identify the pairs of loci a and b, where there is a domain between a and b (equivalently, where the pixel $M^*_{a,b}$ is the corner of a domain). As noted above, it is useful to apply the arrowhead transform to M^*, yielding the arrowhead matrix A. Every domain will produce the two triangles $U_{a,b}$ and $L_{a,b}$ described above. By empirically studying the results of A on a series of domains, we noted the following facts about $U_{a,b}$ and $L_{a,b}$:

1. almost all entries in $U_{a,b}$ are negative, and almost all entries of $L_{a,b}$ are positive.
2. when the sum of the entries in $U_{a,b}$ is subtracted from the sum of the values in $L_{a,b}$, the resulting value is large (relative to a random model)
3. the variance of the entries in $U_{a,b}$ and $L_{a,b}$ were both small (relative to a random model).

These properties were not satisfied when $M^*_{a,b}$ was not a domain corner. We therefore used these three observations as a heuristic to find domain corners. To calculate the corner score for a pixel $M^*_{a,b}$, we first calculate a set of subscores for the corresponding $U_{a,b}$ and $L_{a,b}$: S_{sign}, the sum of the signs of entries in $L_{a,b}$ minus the sum of the signs of the entries in $U_{a,b}$; S_{sum}, the sum of the values of entries in $L_{a,b}$ minus the sum of the values of entries in $U_{a,b}$; and $S_{variance}$, the total variances of both $U_{a,b}$ and $L_{a,b}$. We normalize each of these three subscores by calculating each score for every possible a,b, and then dividing by the maximal value observed. The 'raw corner score' matrix

S' comprises the sum of the three normalized scores for all pixels $M^*_{a,b}$. If $M^*_{a,b}$ is a true domain corner, the value of $S'_{a,b}$ will typically be large.

To identify domain corners using the corner score, we create a filtered version of the matrix S', labeled S, in which we set all pixels whose individual subscores do not pass certain thresholds to zero. These thresholds were determined empirically; we believe most of the genome to be partitioned into domains, but erred on the side of fewer false positives when choosing thresholds. We apply thresholding twice, and in each round choose two thresholds, t_1 and t_2. In the first pass, we look for small, very distinct blocks with low variance ($S_{variance} < 0.2 = t_1$; $Mean(sgn(U_{a,b})) < -0.5 = -t_2$; $Mean(sgn(L_{a,b})) > 0.5 = t_2$). In the second pass, we identify larger blocks ($Mean(sgn(U_{a,b})) < -0.4 = -t_2$; $Mean(sgn(L_{a,b})) > 0.4 = t_2$). These larger blocks are not permitted to contain any of the previously annotated smaller blocks.

When we examine the matrix S, we find that corners of domains appear as blobs of high scoring pixels. To precisely annotate domain corners, we first use MATLAB's connected component algorithm to identify groups of adjacent pixels. The pixel within the connected component whose corner score S is largest is marked as the domain corner. See Figure 4.1C,F.

4.1.4 Dynamic programming for fast calculation

Naively, the above algorithm would require us to calculate all the above noted scores for $U_{a,b}$ and $L_{a,b}$ for all a,b. Thus, the naive running time of the above algorithm is $O(n^4)$, where n is the number of loci in the genome. This makes the algorithm infeasible on large-scale datasets.

However, we developed a dynamic programming implementation of this scheme which requires only $O(n^2)$ operations, which makes the algorithm much more useful in practice.

To create a more practical implementation, we realized that summing entries of a matrix contained in $U_{a,b}$ and $L_{a,b}$ can be thought of as summing the calculations for smaller triangles, plus a sum for the additional row or column. In particular, given the sum for $U_{a,b-1}$, we add the column b sum from rows $(a+b)/2$ to a, and similarly for $L_{a,b-1}$. The additional column and row sums are themselves calculated ahead of time via dynamic programming and then accessed when needed to calculate sums for all possible $U_{a,b}$ and $L_{a,b}$

This approach can be applied very broadly. For instance, the variance of the entries in $U_{a,b}$ and $L_{a,b}$, the score matrix $S_{variance}$, can be calculated using dynamic programming by transforming the problem into a sum, relying on the fact that $Var(X) = E[(X - \mu)^2] = E[X^2] - (E[X])^2$.

By exploiting this method, all the above scores can be calculated using only $O(n^2)$ operations.

4.1.5 Properties of Domains

The Arrowhead algorithm on GM12878 was performed at 5 kb resolution. When we performed the Arrowhead algorithm separately on our primary and our replicate GM12878 maps, we called 7105 and 6082 domains respectively, 5041 of which overlapped. This suggested high reproducibility of our domain annotation algorithm. We then applied the Arrowhead algorithm to our combined GM12878 map; the resulting annotation of 9,274 domains was used for subsequent analyses.

Interaction probability drops at the boundaries of domains: Loci within a domain preferentially form contacts with other loci inside the domain relative to neighboring loci outside the domain. This creates a drop in contacts at the borders of domains that is visually apparent in Hi-C maps. To quantify this drop in contacts, we

assessed the ratio of inter-domain contacts to intra-domain contacts at various distances, d, in our in situ map at 25 kb resolution. To do this, we took all pairs of 25 kb loci (that were separated up to a maximum distance of 475 kb) and split these pairs into two lists: those for which both loci in the pair were in the same domain, and those for which the two loci were not in the same domain (at least one locus had to be annotated inside a domain). We then calculated the mean contact frequency at a given distance for each of the two lists. Figure 4.1G shows the ratio of the two mean contact frequencies as function of distance.

Domains exhibit consistent patterns of histone modifications: To determine how domain structure affected chromatin marks, we first took each of our domains and divided it into 10 bins, where the bin size was a tenth of the size of the domain. For each domain, we then recorded the mean value of the chromatin mark of interest within each of these bins. We also recorded the mean chromatin mark value in the ten bins to the left and to the right of the domain boundaries, where the bin size was set to a fixed size (10 kb). This was the procedure used for the matrices shown in Fig 2B. We repeated this procedure with one additional variation, setting the size of the loci flanking a domain to the size of the loci within the domain. This was the procedure used for the matrices shown in Fig 4.1I-J. Results from both methods were similar. To control for outliers, bins whose mark values were above the 99.9th percentile of all bins over all domains were reduced to the value of the 99.9th percentile.

For a chromatin mark of interest, the above procedures yielded a matrix whose length was the number of domains, and whose width was 30. By calculating the correlation of the columns of this matrix, we obtain a 30x30 correlation matrix that can be computed for any specific chromatin mark. This correlation represents how correlated the chromatin marks are at any two loci, and makes it possible to explore the effects of domain

boundaries. The correlation matrices show that chromatin marks exhibit strong positive correlation within domains, and a sharp drop in correlation at the domain boundaries. This is in marked contrast to the result we obtain when we randomly shuffle our domain list. For random domains, we expect loci near each other in the genome to have correlated chromatin marks; however we do not expect anything special to occur at the random domain boundaries. Indeed, the correlation matrices for chromatin marks near random domains are smooth.

We can also compute the correlation between different chromatin marks at different loci relative to a chromatin domain. We expect repressive marks like H3K27me3 to anticorrelate with active marks such as H3K36me3. This is what we see. The correlation matrices between two chromatin marks also display strong drops at the domain boundaries, in contrast to the random domain correlation matrices. All correlations matrices (for true and random domains), with the bin size kept constant both inside and outside of domains, are shown in Figure 4.1K-L.

Recently, Naughton et al.[103] devised a high-throughput experiment to measure the amount of supercoiling in cells by using biotinylated 4,5,8-trimethylpsoralen (bTMP) as a probe. They measured the amount of bTMP binding on chromosome 11, where enriched binding (relative to input) indicates negative supercoiling and depleted binding indicates positive supercoiling. In Naughton et al. the authors found that 30% of the boundaries of their annotated supercoiling domains fell within 20 kb of topological domain boundaries. Here we examined the bTMP signal/input track across whole domains in chromosome 11, and found that, like many of the epigenetic marks discussed here, there was higher correlation of supercoiling signal between loci located in the same domain than between loci located in different domains (see Figure 4.1J). While the correlation drop at the domain boundaries does not seem as sharp as the drops that

appear in the chromatin modification correlation matrices, it is certainly enriched above random domains, indicating that supercoiling status is related to domain structure. This is consistent with early studies[31,143,153,54] regarding chromatin organization which posited that the genome was folded into distinct 'topologically constrained domains' or 'chromatin domains', each of which was thought to be "a unit of supercoiling, in that its torsional state is independent of the torsional state of the surrounding loops"[54].

We also computed the correlation of chromatin signals at a fixed distance for loci in the same and in different domains, and found for instance, the correlation between H3K36me3 (resp. H3K27me3) signals for two loci 50 kb apart was 0.52 (resp. 0.59) if the loci were in the same domain, but only 0.23 (resp. 0.19) if they were not. For this calculation we took all pairs of 25 kb loci separated by a 50 kb interval and split them into two lists: pairs for which both loci were in the same domain, and pairs for which the two loci were in different domains (at least one locus had to be annotated inside a domain). We then calculated the Spearman correlation of chromatin marks at these pairs of loci, and found all marks are more correlated if they are in the same domain than if they are in different domains (see Figure 4.1I).

We note that, while we see that all marks are more correlated between loci in the same domain than between loci in different domains, the strength of the correlation will depend on many factors. The various chromatin marks differ in terms of how easily they spread, how specifically they bind to punctate versus broad features in the genome, how frequently they appear in a given cell-line, etc. All of these factors will influence the strength of the correlation when averaged over the length of the domain; thus, in order to compare correlation values between any two chromatin marks, care must be taken to account for these differences.

Changes in patterns of long-range contact tend to occur at the boundaries

of domains: When examining Hi-C maps, we noticed that loci within a domain seemed to have the same long-range interaction pattern, while changes in long-range interaction patterns occurred on the boundary between domains. To quantify this, we devised a gradient score, which measured the difference in long-range interaction pattern between all neighboring loci.

For each 25 kb locus, i, along the genome, we calculated a score $G_{i,j}$ at every pixel $M^*_{i,j}$ for all $|i-j| > 10$ Mb and < 40 Mb, where:

$$G_{i,j} = \frac{(A_{i,j} - E_{i,j})^2}{E_{i,j}} + \frac{(B_{i,j} - E_{i,j})^2}{E_{i,j}} \tag{4.1}$$

and:

$$A_{i,j} = \sum_{a=i-4}^{i-1} \sum_{b=j-2}^{j+2} M^*_{a,b} \tag{4.2}$$

$$B_{i,j} = \sum_{a=i+1}^{i+4} \sum_{b=j-2}^{j+2} M^*_{a,b} \tag{4.3}$$

$$E_{i,j} = \frac{A_{i,j} + B_{i,j}}{2} \tag{4.4}$$

Our final gradient score for every locus i, G_i, was the sum of all ${}_{G}i,j$ for all $|i-j| > 10$ Mb and < 40 Mb. We then examined the distribution of G at bins inside domains and at domain boundaries, and then repeated this procedure with domains defined by a random shuffle of our domain list. Values of G were higher at the boundaries of true domains and were depleted within domains as compared to our randomly shuffled domain list (see Figure 4.1H), indicating that changes in long range interaction patterns tend to occur at domain boundaries.

4.2 Clustering Methodology

4.2.1 Previous approaches for clustering contact patterns

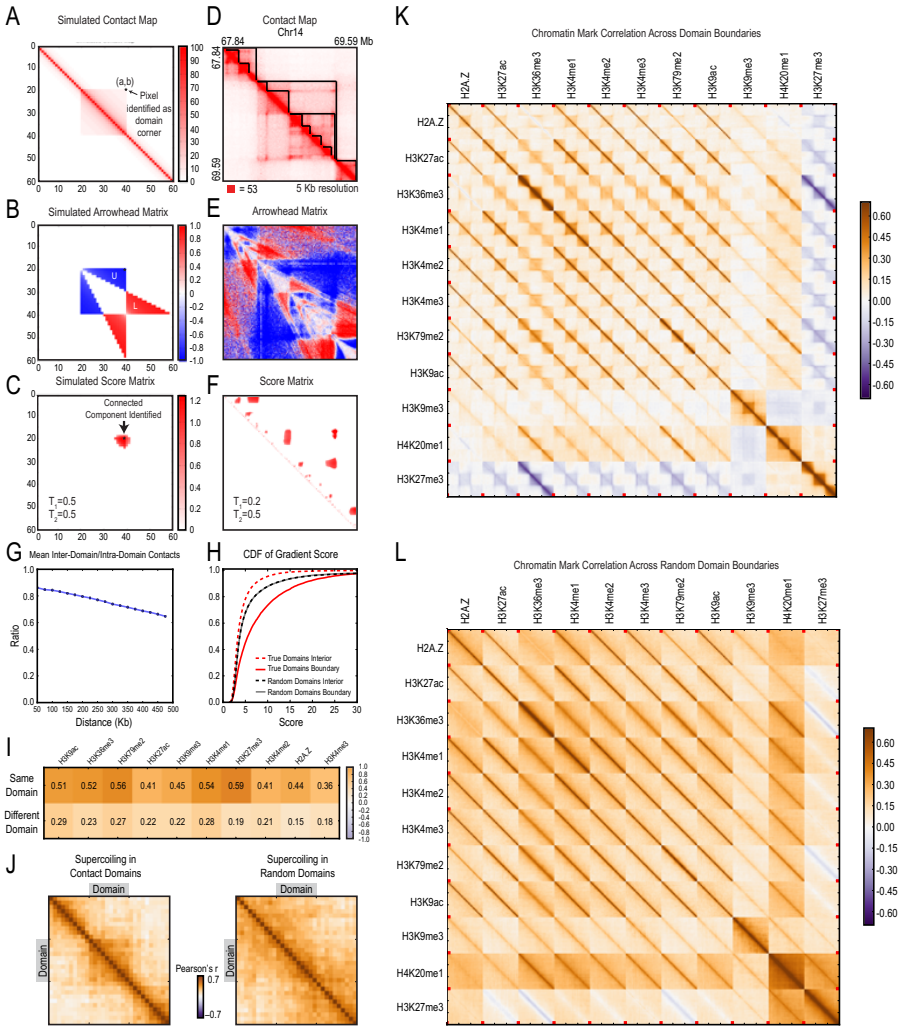
The most common method used for classifying Hi-C patterns is the principal component (PC) approach, which we introduced in[83]. In this approach, each intrachromosomal contact matrix is converted to an observed/expected matrix, and the first principal component of this matrix is used to bifurcate the data into two clusters. When this was performed at 1 Mb resolution on the original Hi-C maps, it was found that one cluster (A) was enriched for open chromatin marks, while the other cluster (B) was enriched for closed chromatin.

When we tried this approach on our higher resolution data, we found that it did not capture all of the patterns we saw, and in fact misclassified some B-type patterns. We therefore wondered if other methods might do a better job of matching the multiple distinct patterns that we observed.

Previous work in the field has studied alternatives and modifications to the 2-compartment model. Yaffe and Tanay[149] performed a 3-pattern clustering of Hi-C data, applying k-means on the interchromosomal contact matrix, and found a third, gene-poor cluster. Imakaev et al.[66] also examined possibilities outside the 2-compartment model; they found that the principal components of contact matrices vary continuously, and caution that classification into two clusters may be incomplete. With our larger maps, we sought to update the classification of contact patterns using the clustering method described below.

Figure 4.1 ***(following page)*****:** (A-C) We apply the Arrowhead algorithm to a simulated contact map (A); the result is the Arrowhead matrix (B), where the domain has been transformed into a pair of triangles of opposite signs, named U and L. Using a heuristic scoring algorithm (with thresholds T1=0.5 and T2=0.5) we obtain the Score matrix (C), where the pixels located in the corner of the contact matrix appear as a patch of high values. The pixel with the maximum score in this patch is annotated as the domain corner, which in this case is exactly the corner of the simulated domain. (D-F) Same as on the left, except we apply the Arrowhead algorithm to a 1.75 MB region in chromosome 14 in our in situ GM12878 map (D). Domains are transformed into arrowheads (or pairs of red and blue triangles) in the Arrowhead matrix (E). The Score matrix (obtained using T_1=0.2 and T_2=0.5) identifies areas of high corner likelihoods and the maxima are marked as domain corners (F). Our domain calls for this region are shown in black in (D). (G) The ratio of the mean inter-domain contacts to mean intra-domain contacts at various distances; contacts between loci in different domains are depleted, accounting for the drop in contacts we observe at domain boundaries. (H) The cumulative distributions of the long-range gradient score G; higher values of G indicate greater differences in the long-range patterns of neighboring loci. The gradient score is enriched at domain boundaries (solid red), relative to random domain boundaries (solid grey), implying that pattern switches happen at domain boundaries. The interior of domains (dashed red) is depleted for long-range pattern changes relative to random domain interiors (dashed grey). (I) Spearman correlation coefficient of various chromatin marks for loci separated by 50kb. Loci in the same domain (top row) show higher correlations for chromatin marks than loci in different domains (bottom row). (J) Correlation of supercoiling within domains, using data from Naughton, et al[103]. Loci within the same domain show higher correlations than loci in different domains, in contrast to the smooth correlation matrix (right) calculated for randomly placed domains. (K) Correlation matrices for 11 chromatin marks in relation to domain boundaries in our in situ GM12878 map. Domains were divided into 10 equally sized bins, and the bins were extended 10 to the left and to the right of the domain borders (with the bin size kept the same both inside and outside of the domains). For each mark, we calculated its 30 by 30 correlation matrix in these regions (the middle 10 by 10 values of this matrix indicate correlations of the mark at loci within the domain). Marks are correlated within domains, but the correlation sharply drops at domain borders. We also show the correlation matrices for all pairs of epigenetic marks in relation to domain boundaries, and find strong correlation or anticorrelation of marks within domains that drops at domain boundaries. We combine all of the correlation matrices into one large image. Domain boundaries are marked by black tick marks; red tick marks separate the different epigenetic marks. (J) Same as (K), except we show the correlation matrix for 11 chromatin marks in relation to random domains (instead of true domains), in our in situ GM12878 map. No sharp changes occur at domain boundaries.

Figure 4.1: (continued)



4.2.2 Clustering Algorithm

To cluster loci based on long-range contact patterns, we constructed a 100 kb resolution contact matrix C comprising a subset of the interchromosomal contact data. 100 kb loci on odd chromosomes appeared on the rows, and 100 kb loci from the even chromosomes appeared on the columns. The total length in base pairs of these two groups is roughly equal. (Chromosome X was excluded due to the differences in interaction pattern seen for the active and inactive homologs.) Thus $C_{i,j}$ represents the number of normalized contacts between the i-th locus on the odd chromosomes and the j-th locus on the even chromosomes. Genome-wide KR was used for normalization. Rows and columns for which more than 30% of the entries were either undefined or zeros were removed from the matrix. These bins were excluded from all further analyses involving the cluster tracks unless otherwise noted. We then took the logarithm of each entry in the C matrix.

To cluster loci on the odd chromosomes, we applied the z-score function from Python's scipy library to each row of C. We used the resulting matrix as input to the scikit-learn library's unsupervised Gaussian hidden Markov model clustering algorithm (GaussianHMM)[112]. We set the covariance type to diagonal and allowed 1000 iterations.

To perform clustering on loci located on the even chromosomes, we began by transposing the matrix C and then performed all steps exactly as they were performed for the clustering of odd chromosomes.

We found that each of the clusters on the odd chromosomes preferentially interacted with one of the clusters on the even chromosomes (Fig. 4.2D). This defined a one-to-one mapping between the odd and even cluster annotations.

The result of a clustering algorithm typically depends on the choice of a parameter,

k, which determines the number of clusters to be identified. We report the results of clustering using k=5 clusters; however, we also performed clustering using all values between k=2 and k=14 clusters as input. The Akaike Information Criterion and Bayes Information Criterion for the different cluster results clearly ruled out a value of k=2, and suggested a value of k between 4 and 8. Our final use of k=5 was based on this finding as well as careful examination of the data to determine how many clusters were necessary to explain the patterns that could be visually discerned. We found that, for k=5, the clusters corresponded to visually distinct patterns; this was no longer true if we increased k beyond 5. Nonetheless, it is possible that there are additional clusters that our algorithm could not identify; our results should be considered a lower bound on the number of subcompartments rather than an exact determination. Clustering via k-means and hierarchical clustering yielded similar results (Fig. 4.2B).

4.2.3 Creating an A/B Pattern Annotation

To define the pattern of a cluster, we use the first derivative of interactions along the linear genome. More precisely, for each locus i on an odd chromosome, we obtain its 1-dimensional interchromosomal interaction vector with all of the even chromosomes, C_i, and then calculate $d_i(j) = [C_i(j) - C_i(j-1)]$, where j and $j-1$ are adjacent loci on an even chromosome. The intuition for using such a measure is based in how we expect the interaction vector C_i for a given locus to change (or switch) when it exits one cluster and enters another cluster. When locus i is interacting with a stretch of loci (on an even chromosome) that are all in the same cluster, the derivative is close to zero, as the amount of interaction, whether high or low, does not change. However, at the border between two different clusters, when j and $j-1$ are in different clusters on the other chromosome, we expect $|d_i(j)|$ to be large. It is these switches that we use to

determine cluster similarity. Using the derivative as a measure of pattern similarity is a simple way to account for the one-dimensional nature of the polymer. (This is akin to measures in finance that correlate returns of prices to identify similarities between stocks.)

To use this measure, we first create the difference matrix D by taking the difference between every adjacent pair of columns in the odd/even interchromosomal matrix C mentioned above. We then calculate a mean vector for each of the clusters (on the odd chromosomes) by averaging the rows of D for loci within the same cluster. Next, we examine the Spearman correlation of these mean derivative vectors for different pairs of clusters.

When examining the correlation matrix of the patterns, we found that the 5 patterns separate into two groups, with A1 and A2 in one group, and B1, B2 and B3 in the other. Patterns in a group correlate with each other and anticorrelate with patterns of the other group (Fig. 4.2E). These correlations were confirmed by visually examining the five patterns. The first group of patterns correlated with the A compartment and the second group correlated with the B compartment.

V.a.4. An additional cluster on Chromosome 19: Careful visual inspection of the contact map for chromosome 19 revealed an additional, distinctive pattern which was visually apparent on a small set of loci but which was not found by the clustering algorithm. We suspected that this was because these loci occupy only 11 Mb, or 0.3% of the genome. We therefore created a matrix containing all interchromosomal contact data for chromosome 19, excluding the X chromosome. We processed the data as above, calculating the log of each entry and a z-score. (Rows and columns for which more than 50% of the entries were either undefined or zeros were removed from the matrix prior to normalization.) Finally, we used the same Gaussian HMM algorithm as before, using

k=5.

One of the clusters the algorithm returned (labeled 19*) corresponded precisely to loci exhibiting the sixth pattern that we had noticed on visually inspecting the matrix. The other four chromosome 19-specific clusters (labeled 19-1 through 19-4) needed to be matched up to clusters found genome-wide in order to create a single classification. We did this by examining the frequency of interaction between the 19-specific clusters and the clusters previously labeled in the interchromosomal contact matrix. Cluster 19-1 interacted most frequently with loci in B2; Cluster 19-2 interacted most frequently with cluster B1; and Cluster 19-3 interacted most frequently with cluster A1. Cluster 19-4 interacted most frequently with A1, and next-most-frequently with B1. Since A1 had already been assigned to 19-3, and 19-4 exhibited H3K27me3 enrichment and H3K36me3 depletion, we labeled the 3.3 Mb that fell in 19-4 as B1.

To determine whether the loci in cluster 19* were in the A compartment or the B compartment, we compared the derivative of its intrachromosomal contact pattern to the derivatives of the subcompartments in the A and B patterns, which revealed a stronger correlation to compartment B patterns (Fig. 4.2F). This was confirmed by visual inspection (see Figure 2F), and led us to label this cluster B4. It is worth noting that, interchromosomally, the B4 cluster pattern's derivative is more correlated with compartment A-type derivatives, and that loci exhibiting the B4 pattern also tend to possess both activating and repressive chromatin marks. These observations suggest that subcompartment B4 may be difficult to classify as either A or B, as it appears to have some degree of affinity for both.

4.2.4 Properties of Subcompartment Clusters

Clusters display distinct chromosomal and size distributions: There were differences in the distribution of clusters among individual chromosomes. For example, the larger chromosomes had much more B3 than B2, while the opposite was true for the smaller chromosomes (Fig. 4.2G). The total number of megabases of the genome covered by each of the clusters is given in Fig. 4.2C. Many adjacent 100 kb loci belong to the same cluster. We call a stretch of contiguous loci belonging to the same cluster a "cluster interval". The median size of cluster intervals is 300 kb. The mean and median size of cluster intervals for the different clusters is given in Fig. 4.2C.

Clusters display unique patterns of epigenetic modifications: After clustering the data, we sought to determine enrichment of epigenetic signal tracks in the different clusters. To do so, we first binned the signals into 100 kb bins (taking the mean in each bin). For the enrichment analysis (Figure 2D), we calculated the median value of the signal track in bins within the cluster of interest and divided that by the median value of the signal track across all bins. To determine how the signal tracks correlated with the clusters, we calculated the Spearman correlation coefficient between the binned signal track and a pseudo cluster track, where the pseudo cluster track had 1s at each 100 kb locus that belonged to the cluster of interest and -1s at all other loci (Fig. 4.2I). We also performed the Wilcoxon rank-sum test, comparing the values of the mark in loci within the cluster of interest to the values of the mark in all loci of the remaining clusters, and the results were similar.

For the B4 cluster, we observed a simultaneous enrichment of H3K36me3 and H3K9me3, a seemingly paradoxical combination of activating and repressive marks. Additionally, we observed an enrichment of H3K9me3 over the A2 compartment when compared to the

A1 compartment (both A1 and A2 contain active, open chromatin). In order to confirm that these results did not occur as a consequence of inappropriate crosstalk because of a faulty H3K9me3 antibody used in the ENCODE experiment we utilized, we repeated the H3K9me3 ChIP-Seq experiment seven times with three different antibodies from three different companies (see Section I.b). In all cases, the simultaneous enrichment of H3K9me3 and H3K36me3 over the B4 compartment and the enrichment of H3K9me3 on the A2 compartment compared to the A1 compartment was reproduced. In fact the simultaneous enrichment of H3K36me3 and H3K9me3 over KRAB-ZNF genes (which make up most of the B4 compartment) has been noted several times previously and reproduced by many groups[142,11].

In Chapter 2, we calculated enrichment of Nucleolar Associated Domains (NADs), which is encoded in a binary track, by taking the ratio of the length (in base pairs) of NADs found in the cluster and the length (in base pairs) of NADs found in the cluster when it was randomly permuted in the genome[110]. For the supplemental figure, we examined the correlation with a continuous Nucleolus association track from[110]. Enrichment for nuclear periphery was evaluated by association of the clusters with the results of a Chip-Seq experiment from[90] using lamin A/C, a protein known to localize at the nuclear periphery.

Pericentromeric chromatin was defined as the 2 Mb before and after each centromere, where the locations of the centromeres were given by the hg19 consensus annotation (`http://genome.ucsc.edu/cgi-bin/hgTables`). Of the 54.2 Mb of pericentromeric chromatin that was annotated by our clustering algorithm, 33.8 Mb of it (62.3%) is located within B2, a 3.8-fold increase from what would be expected at random. B1 contains 5.9 Mb, a 4% increase over expected. The other clusters are depleted for pericentromeric chromatin.

We also examined GM12878 Repli-Seq data, and found A1 and A2 were enriched for early-replicating chromatin, and B2 and B3 enriched for late-replicating chromatin, in agreement with[121] and[58]. More specifically, we found that A1 and A2 both begin to replicate in the G1 phase; however, A1 tended to finish replicating by the S1 phase while A2 continued to replicate through the S2 phase. B2 and B3 do not begin replicating until the S3 phase and replicate primarily in the S4 and G2 phases. B1 begins in the G1 phase, but primarily replicates in the S1, S2, and S3 phases (Figure 2D).

4.3 Aggregate Peak Analysis

Aggregate Peak Analysis (APA) is a method that allows us to test the aggregate enrichment of an entire set of putative two-dimensional peaks, as opposed to verifying individual peaks one-by-one. This method is especially useful for checking a set of peak calls on a low-resolution Hi-C map, where individual peaks may be impossible to discern but where the aggregate signal from the full peak set should be detectable if the peak set is reliable and the Hi-C map was the result of a successful experiment in the same cell type. Notably, we have not found an instance in which one of the in situ Hi-C peak lists reported here failed to be validated, in aggregate, when examined with respect to a successful Hi-C experiment in the same cell type.

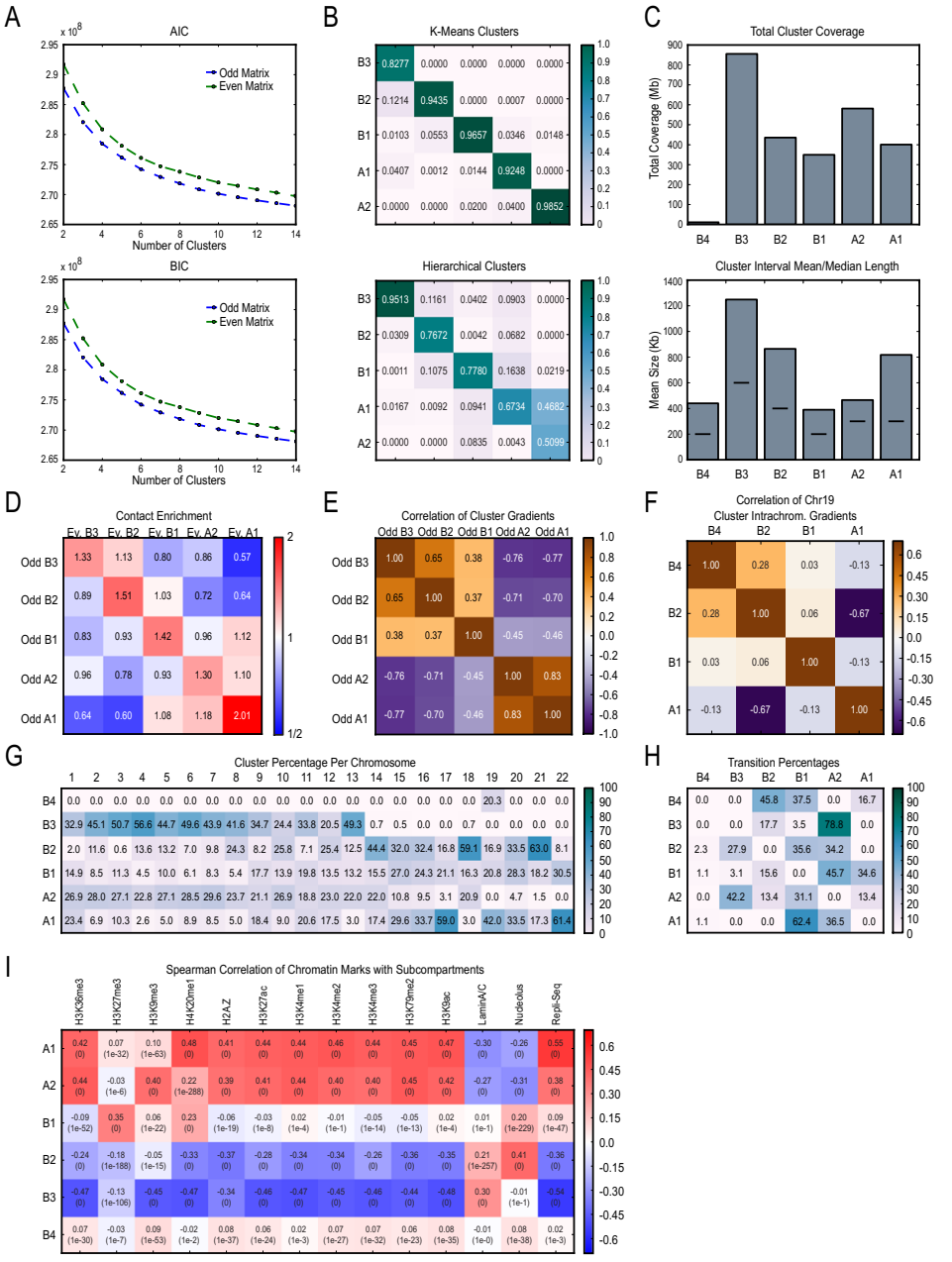
4.3.1 Standard APA analysis

APA quantifies the enrichment of a peak set in aggregate by plotting the sum of a series of submatrices derived from a contact matrix. These submatrices are chosen so that each one surrounds a single putative peak pixel (note that for intrachromosomal pixels, we always choose the pixel in the upper-right half of the matrix). In the resulting APA plot, the total number of contacts that lie within the peak pixel set is shown at the center; the entry immediately to the right of center corresponds to the total number of contacts in the pixel set obtained by shifting the peak set 10 kb to the right; the entry two positions above center corresponds to an upward shift of 20 kb; and so on. Focal enrichment across the peak set in aggregate manifests as larger values at the center of the APA plot.

To perform APA, a resolution and window size is chosen. In Figure 3, we have chosen 10 kb resolution and a window of $\pm$ 100 kb (10 pixels) about each putative

Figure 4.2 *(following page)*: (A-B) We show the (A) Akaike Information Criterion and the Bayes Information Criterion (which differed only slightly) for the HMM clustering as the number of clusters is varied. While no sharp transition is seen, the curve suggests that choosing k between 4 and 8 is appropriate. (B) When we clustered our interchromosomal matrix using K-means or hierarchical clustering instead of an HMM, the results were highly similar. Here we show the percentage of each of the 5 K-means/Hierarchical clusters that overlapped the HMM clusters (comparing the even K-means/hierarchical clustering to the even HMM clusters). (C) The total number of base pairs assigned to each of the clusters (cluster coverage) and the mean size of a contiguous cluster group (adjacent loci that share the same cluster assignment). Median sizes are shown with horizontal lines. (D) Contact enrichment between the two k=5 clustering assignments (clustering was performed separately on odd and even chromosomes). The mean number of contacts for each pair of clusters is divided by the expected number of contacts (calculated from the average of 100 shuffled clustering annotations). Each cluster on the odd chromosomes preferentially interacts with only one cluster on the even chromosomes (and vice versa); we merged each strongly interacting pair of clusters into a single cluster. (Data was normalized using Interchromosomal KR.) (E-F) Cluster similarity, determined by the Spearman correlation of the one-dimensional derivative vectors. In the genomewide annotation (E), B1, B2, and B3 have similar patterns, as their derivative vectors all correlate with each other, and A1 and A2 have similar pattern, as they correlate with each other. A1 and A2 anticorrelate with B1, B2, and B3. (F) For Chr19, the derivative of the Chr 19 O/E matrix was calculated, and then collapsed onto a matrix of four rows, where each row represented a cluster and contained the mean derivative for that cluster. Pearson correlations of these derivatives were calculated for each pair of clusters. B4 is more closely correlated with B2 and B1 than it is with A1, suggesting a B-type pattern. (G) The percentage breakdown by chromosome for each of the clusters; various patterns have preferential locations in the human genome. (H) The percentage of transitions that occur from a cluster (indexed by the rows) to another cluster (indexed by the columns). Rows sum to 100. For example, A1 most frequently transitions to B1. (I) Spearman correlation coefficients (p-values in parentheses) for various epigenetic marks and the 6 clusters. Each cluster displays a unique combination of marks. Results using the Wilcoxon rank-sum test were similar.

Figure 4.2: (continued)



peak. The APA resolution determines the resolution at which the contact matrix of a Hi-C map is generated. Note that, in APA analyses, contact matrices will often be generated for a given Hi-C map at resolutions vastly higher than the resolution at which the map would usually be examined. For instance, we generated contact matrices for our 2009 Hi-C maps[83] at 10 kb resolution, despite the fact that the map resolution of these maps is orders of magnitude larger. (The aggregation process makes it possible to resolve features at much higher resolutions than would ordinarily be possible with a map, producing a "super-resolution" image.)

We center a submatrix at each peak in the target peak set at the chosen resolution. The width of the matrix is the window size above. For peak calls that span an area larger or smaller than one pixel in the chosen resolution, we choose the center of the peak call as the center of the matrix. Only one submatrix is created per peak call, even if the peak call extends to multiple pixels. If the center pixel of multiple peak calls falls into the same pixel, we use that submatrix only once. To avoid strong distance effects, we only examine peak calls where the peak loci are separated by more than a minimum threshold t.

For APA performed at 10 kb resolution with a window of $\pm$ 100 kb, $t = 300$ kb. For APA performed at 5 kb resolution with a window of $\pm$ 25 kb, $t = 100$ kb.

Submatrices for the peaks are taken from the normalized (intrachromosomal KR corrected) Hi-C maps. These submatrices are then summed (entry-wise), obtaining an APA matrix in which the center pixel represents the sum of the number of reads in the entire target loop set. For the MiSeq datasets, a KR correcttion is not available at 10 kb due to the sparsity of the data, so the KR correction factors for all the maps are approximated by first calculating them for a coarser resolution.

To determine if the center pixel of the APA plot is focally enriched, we calculate the

APA score, which is the ratio of the number of reads in the center bin to the average number of reads in the lower-left corner of the APA matrix. We define the lower left bins for 10 kb resolution and a ± 100 kb window size as the bins lying in the bottom-left 6 x 6 square of the matrix. For 5 kb resolution and a ± 25 kb window we define the lower left bins as those lying in the lower left 3 x 3 square of the matrix. We chose to use this particular score because it is very simple to understand and calculate, and because it corresponds to the widely-accepted notion of a loop: in order for the APA score to be above 1, the number of contacts between a typical pair of loop anchors in the peak set must be higher than the number of contacts between intervening pairs of loci.

To calculate a p-value for this score, we calculate the z-score which compares the central bin to the set of bins in the lower left window defined above. The z-score is then converted into a p-value (1-sided).

The color scale in all APA plots is set as follows. The minimum of the color range is 0. The maximum is 5 x UR, where UR is the mean value of the bins in the upper-right corner of the matrix. The upper-right corner of the 10 kb resolution APA plots is a 6 x 6 window (or 3 x 3 for 5 kb resolution APA plots).

4.3.2 Control peak sets

We created control peak sets to investigate the effect of domains on APA. To generate these control peak sets, we randomly choose pixels inside of the GM12878 domains that we annotated using the Arrowhead algorithm. The control peak sets did not display APA enrichment.

4.3.3 Advanced APA analyses

Although the paper focuses on the most straightforward form of APA analysis, which is described above, we also explored a number of more sophisticated methods.

A simple variant on standard APA is "normalized" APA. In normalized APA, each submatrix is "normalized" by dividing all entries of the submatrix by the mean value of the submatrix, such that the mean value of the entries of the resulting submatrix is 1. This ensures that short-range loops do not exert a disproportionate influence on the results of the analysis.

There are also more sophisticated APA scores. For instance, the standard APA score is the ratio of the central bin to average of the lower-left corner. One can check this ratio for all four corners of the matrix. The upper-right is a less useful control: even a random set of pixels should be enriched relative to the upper-right corner of the matrix, due purely to distance effects. The upper-left and lower-right corners comprise pairs of loci that are close to the same distance as the peak loci, so the ratio of the central bin to these corners should be close to 1 for control peak sets. However, enrichment relative to these corners may result from the fact that pixels in the peak set tend to be in domains, rather than at true peaks.

One can also examine the peak to mean value (the ratio of the center bin to the mean of the rest of the matrix). Because the number of contacts varies as a function of distance, this value is less informative. In Fig. 4.3A-B we show examples of these other measures.

We can also perform a more formal measure of enrichment by using the paired Student's T-test. For this, we calculate the one-sided t-statistic, measuring enrichment between the central peak bin and every other bin in the matrix. Values above 2, with

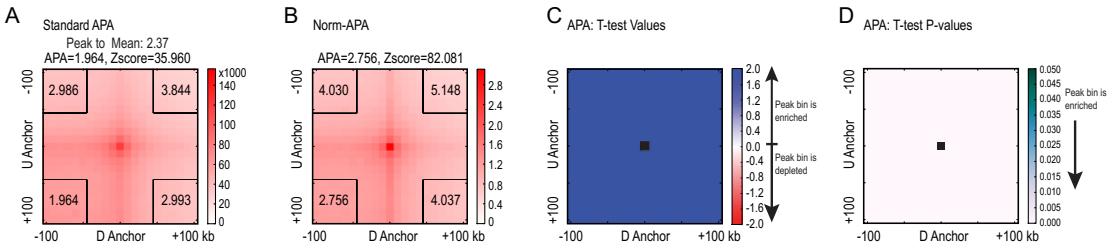


Figure 4.3: (Additional APA measures. (A) Standard APA of our in situ GM12878 peak list examined on our GM12878 dilution map. The APA score is the ratio of the number of contacts in the central bin to the mean number of contacts in the lower-left corner (outlined by the black box). The ratio of the central bin to the other three corners is also shown inside each corner, though we do not use these scores in this paper. APA scores above 1 signify that the peak bin is enriched relative to the bins inside the corner. The ratio of the central bin to the mean of the remaining matrix, along with the z-score of the central bin, using the mean and standard deviation of the lower-left corner, are also shown. Z-scores above 1.64 indicate enrichment ($p<.05$). (B) Normalized APA of our in situ GM12878 peak list examined on our GM12878 dilution map. In Normalized APA, each submatrix is first normalized before being added to the aggregate matrix; normalization is performed by dividing each entry by the mean of the submatrix. The final Normalized APA matrix is the average of all of these submatrices. The maximum color scale in all APA plots is set to five times the mean value in the upper right corner of the matrix. (C-D) We show a more formal measure of aggregate peak enrichment of our in situ GM12878 peak list in our GM12878 dilution map. We use the one-sided paired Student t-test, measuring enrichment between the set of all values in the central peak bin (across all peaks) and the set of values in every other bin in the matrix. T-statistics are plotted in (C), with the corresponding p-values shown in (D). We find that our peak set is statistically enriched ($p<0.05$) relative to every other bin in the matrix. (No multiple hypothesis correction was performed on the p-values).

low p-values, indicate that the central bin is enriched relative to the bin of interest; values below 2 indicate depletion. Fig. 4.3C-D show examples of these matrices.

5

Deletion of DXZ4 on the human inactive X chromosome alters higher-order genome architecture

5.1 Introduction

During interphase, the inactive X chromosome (Xi) is largely transcriptionally silent[22] and adopts an unusual 3D configuration[104,137] known as the Barr body[10]. Despite the importance of X inactivation[131], little is known about this 3D conformation. We recently showed that in humans, Xi exhibits three structural features, two of which are

not shared by other chromosomes. First, like the chromosomes of many species, Xi forms compartments. Second, Xi is partitioned into two huge intervals, called superdomains, such that pairs of loci in the same superdomain tend to co-localize. The boundary between the superdomains lies near DXZ4[119], a macrosatellite repeat[51,126,140] whose Xi allele extensively binds the protein CTCF[24]. Third, Xi exhibits extremely large loops, up to 77 Mb long, called superloops. DXZ4 lies at the anchor of several superloops[63,119]. Here, we combine 3D mapping, microscopy, and genome editing to study the structure of Xi, focusing on the role of DXZ4. We show that superloops and superdomains are conserved across eutherian mammals. By analyzing ligation events involving three or more loci, we demonstrate that DXZ4 and other superloop anchors tend to co-locate simultaneously. Finally, we show that deleting DXZ4 on Xi leads to the disappearance of superdomains and superloops, changes in compartmentalization patterns, and changes in the distribution of chromatin marks. Thus, DXZ4 is essential for proper Xi packaging.

In previous work describing the structure of Xi, we used DNA polymorphisms to assign in situ Hi-C reads to specific chromosomal homologs in order to create a diploid Hi-C map for human GM12878 cells[119]. We showed that Xi has a distinctive superstructure consisting of superdomains and superloops (that is, unusually large domains and loops, spanning dozens of megabases). DXZ4 is situated at the boundary of the superdomains. It also lies at the anchor of superloops to XIST, FIRRE, LOC550643[63] (which we here dub the Inactive-X CTCF-binding Contact Element, ICCE), and a previously uncharacterized locus at ChrX:75350000-75400000 (which we refer to as x75).

We also observed that this Xi superstructure co-exists with compartmentalization. Compartmentalization refers to the fact that many genomes are partitioned into intervals belonging to a handful of types (called compartments), such that loci in intervals of the same compartment tend to co-localize during interphase. In a Hi-C contact

map, these intervals (typically 300 kb long[119]) manifest as square domains of enhanced contact frequency along the diagonal, and the compartments give rise to a 'plaid' appearance. The compartments tend to be associated with different patterns of chromatin marks[83,119].

5.2 Results

In this paper, we use a number of approaches to study the superstructure of Xi and the role of DXZ4.

First, we explored whether the Xi superstructure is conserved among mammals by applying Hi-C to cells from mouse and rhesus macaque. For mouse, we studied the female Patski cell line[85]. Because the cell line was derived from an inter-species hybrid cross [Mus musculus x Mus spretus], it has a high heterozygosity rate (1 in 64 bases) that facilitates assignment of a large fraction of alignable reads (49%) to particular chromosome homologs. Superdomains were apparent in the maps of Xi, but absent from maps of Xa (Figure 5.1A). Notably, despite extensive rearrangement of the murine X chromosome relative to its human counterpart, the boundary between the superdomains again occurred at Dxz4 (chrX: 75.7-75.8 Mb, mm10)[63]. (While this manuscript was in preparation, this point was reported by multiple groups[39,94].) As in human, we also observed superloops between Dxz4 and Firre, Dxz4 and Xist, and Dxz4 and x75 (Figure 5.1B). There was one notable difference, however, between the human and murine Xi maps. Compartment structure was almost entirely absent in the murine Xi: the Hi-C map of Xi did not have a plaid appearance and square domains were rarely seen along its diagonal. The absence of compartmentalization may be a feature of mice in general or may be the result of the cell lines chosen in this and our prior study.

In situ Hi-C in female fibroblasts derived from rhesus macaque (GM08312) similarly

revealed superdomains (Figure 5.1A), with their boundary at the macaque DXZ4 locus (chrX: 114.2-114.3 Mb, rheMac2)[92], as well as a prominent superloop linking the macaque orthologs of DXZ4 and FIRRE (Figure 5.1B).

Taken together, the contact maps of human, rhesus macaque, and mouse suggest that Xi superstructure — including both superdomains and superloops — is widely conserved across eutherian mammals. Because DXZ4's genomic context differs greatly in mouse and human, these findings support a functional role for DXZ4 in determining the placement of superloop anchors and of the superdomain boundary.

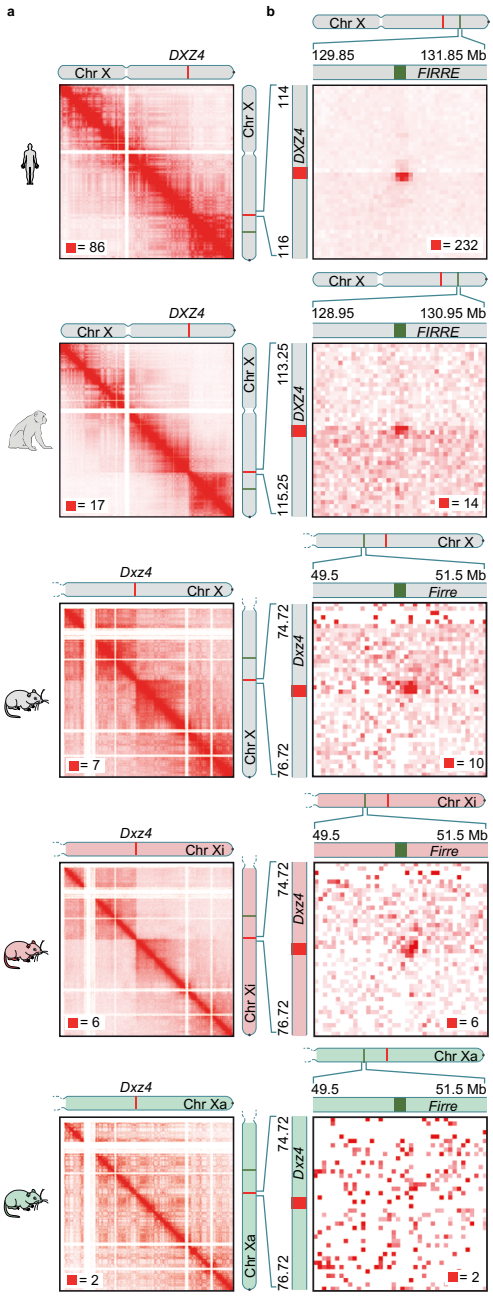
Second, we sought to characterize the structure formed by DXZ4 and other superloop anchors in greater detail. Specifically, we wondered whether superloops tend to occur simultaneously in the same cell, which would suggest the existence of a spatial hub on Xi at which all superloop anchors co-locate, or whether they occur in different cells. To explore this question, we re-examined our recently published Hi-C dataset, specifically the 21 billion reads that were derived from female human cell lines[119]. Although most Hi-C reads bring together two nearby fragments, the procedure can sometimes result in the ligation of three or more nearby fragments in a nucleus. Such events make it possible to examine higher-order spatial relationships within a cell that cannot be interrogated from pairwise contacts. We searched the Hi-C data for reads containing sequences from three or more genomic loci. Although such reads were rare (1 in 2056), the large size of the dataset enabled us to find over 10 million unique 'triples', each exhibiting high-quality alignments (MAPQ-10) to three widely separated loci (all pairwise distances >20 kb) on the same chromosome. The data also contain about 4.6 million unique quadruples and 892 unique quintuples.

Because each triple (*A*-*B*-*C*) implies the proximity of three pairs of loci (*A*-*B*, *A*-*C*, *B*-*C*), we assessed the quality of the triple contacts by creating matrices of the implied pairwise contacts. These matrices closely resembled matrices derived from ordinary, pairwise Hi-C data (Pearson's r=0.96 at 2.5 Mb), suggesting that the higher-order contacts observed using in situ Hi-C reflect genuine patterns of nuclear proximity.

We also developed a modified in situ Hi-C protocol using a restriction enzyme that digests chromatin into much finer fragments, designed to increase the proportion of reads containing three or more nearby fragments. The procedure, dubbed COLA (COncatemer Ligation Assay) uses CviJI, which cleaves RGCY sites between the G and (C) producing short (average of 64 bp) blunt-end fragments that are ligated in situ (Fig-

Figure 5.1 *(following page)*: Xi superstructure is conserved across human, rhesus macaque, and mouse. (A) Superdomains on Xi are conserved across human, rhesus macaque, and mouse. The boundary of the superdomains lies at DXZ4 and its orthologs. In diploid Hi-C maps of mouse, the superdomain is only seen on Xi. (Resolution: 100 kb.) For all contact maps, the color scale of each map goes from 0 (white) to red, whose value is given by the red square in each map. (A) A superloop forms between DXZ4 and FIRRE in human. Superloops are present at orthologous positions in rhesus macaque and mouse. In diploid Hi-C maps of mouse, the superloop is only seen on Xi. (Resolution: 50 kb.)

Figure 5.1: (continued)



ure 5.2A). We sequenced 276M paired-end reads from a COLA library from human GM12878 lymphoblastoid cells. While the pairwise contact matrix closely resembled the one obtained from ordinary in situ Hi-C experiments (Pearson's r=0.82 at 2.5 Mb), the frequency of triples was 13 times as high (1 in 171). (The frequency was 43-fold higher than in dilution Hi-C experiments.) The COLA maps of GM12878 contributed an additional 2,272,943 unique triples, 251,048 unique quadruples, and 1447 unique quintuples.

Such higher order contacts are naturally represented and visualized as an n-dimensional matrix, or tensor. Because the number of entries in this tensor is proportional to the size of the interrogated genome raised to the nth power, the tensor is extremely sparse. Features of chromosome organization manifest inside this tensor as n-dimensional shapes. For instance, the bright diagonal seen in 2-dimensional contact matrices naturally manifests as an n-dimensional hyperstar (Figure 5.2B).

Because the tensor was so sparse, we did not expect to see triples indicating simultaneous occurrence for pairs of ordinary loops, but reasoned that we might see them for DXZ4-FIRRE and ICCE-DXZ4 superloops, because these anchors are so large (up to 300 kb) and thus produce many more fragments. Indeed, we found that the tensor contained a cluster of 4 triples corresponding to the triad of loci ICCE-DXZ4-FIRRE. (Because all three loci are tandem repeats, we included triples that did not align uniquely so long as all of the possible alignments fell in the same locus.) Given the extreme sparsity of the contact tensor, the probability of even a single triple occurring at these three loci at random is <2% (based on locus size and pairwise distance between loci.) The presence of four triples represents an enrichment of 300-fold over random expectation and is extremely unlikely (p=1.4 x 10-9; Figure 5.2C,D). The ICCE-DXZ4-FIRRE voxel was also very strongly enriched with respect to its local, 3D neighborhood (Figure 5.3A,B).

Similarly, we noticed a cluster of four triples at the triad ICCE-x75-DXZ4 (a 154-fold enrichment, p=1.8 x 10-8) (Figure 5.2C,D).

To obtain independent evidence that DXZ4, FIRRE, and ICCE simultaneously co-localize on Xi, we used 3D-FISH. When we examined X chromosomes in male lymphoblastoid cells (GM06992), we did not find a single case in which all three loci co-localized (out of 219 chromosomes examined). For X chromosomes in the female GM12878 cell line, we observed co-localization of all three loci on 17 of 222 X chromosomes (7.7%; p = <0.0001, Fisher's exact test). Because females contain both Xa and Xi, this suggests that all 3 loci overlap >15% of the time on Xi (vs. 0% on Xa) (Figure 5.3C). Because typical pairs of nearby loci (<1 Mb) in the same loop tend to overlap less than 25% of the time when probed using 3D-FISH[119], an observed overlap frequency of >15% among three loci spread across a chromosome is extremely high.

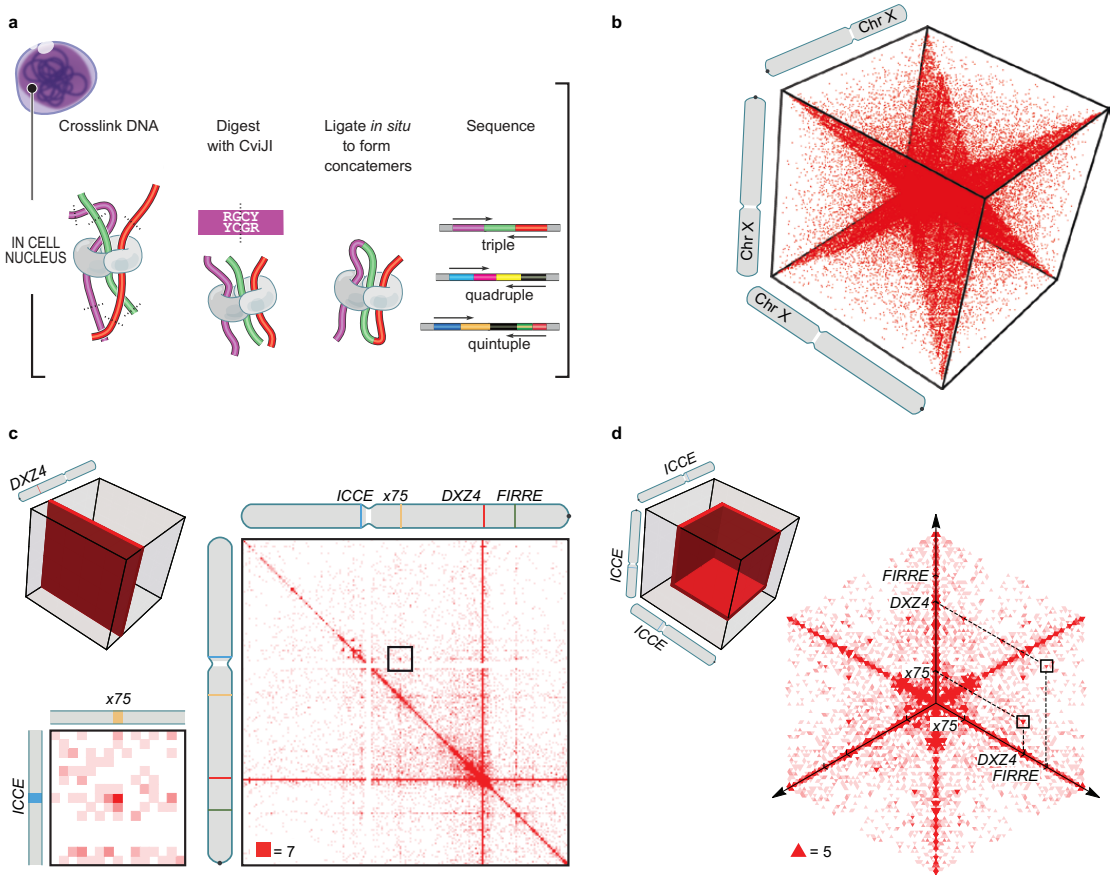


Figure 5.2: Concatemers produced by proximity ligation indicate simultaneous co-location of three or more loci. (A) In COLA, concatemers spanning three or more fragments are created by cutting chromatin using CviJI, followed by DNA-DNA proximity ligation in intact nuclei. (A) Contact triples visualized as a 3D contact tensor. Broad patterns of contacts can be revealed by slicing the tensor and examining the results at low resolution. (C) A planar cut of the contact tensor enables the examination of all triples containing DXZ4. Blowup: Enrichment is seen in the plane at x75(-DXZ4)-FIRRE. (Resolution: 800 kb.) (D) A different cut makes it possible to study triples in the vicinity of ICCE. Superloop triples (ICCE-)DXZ4-FIRRE and (ICCE-)x75-FIRRE are highlighted. (Resolution: 2 Mb.)

Taken together, these findings suggest that the DXZ4, FIRRE, and ICCE loci on Xi tend to co-locate at a single spatial position, i.e., that they form a hub.

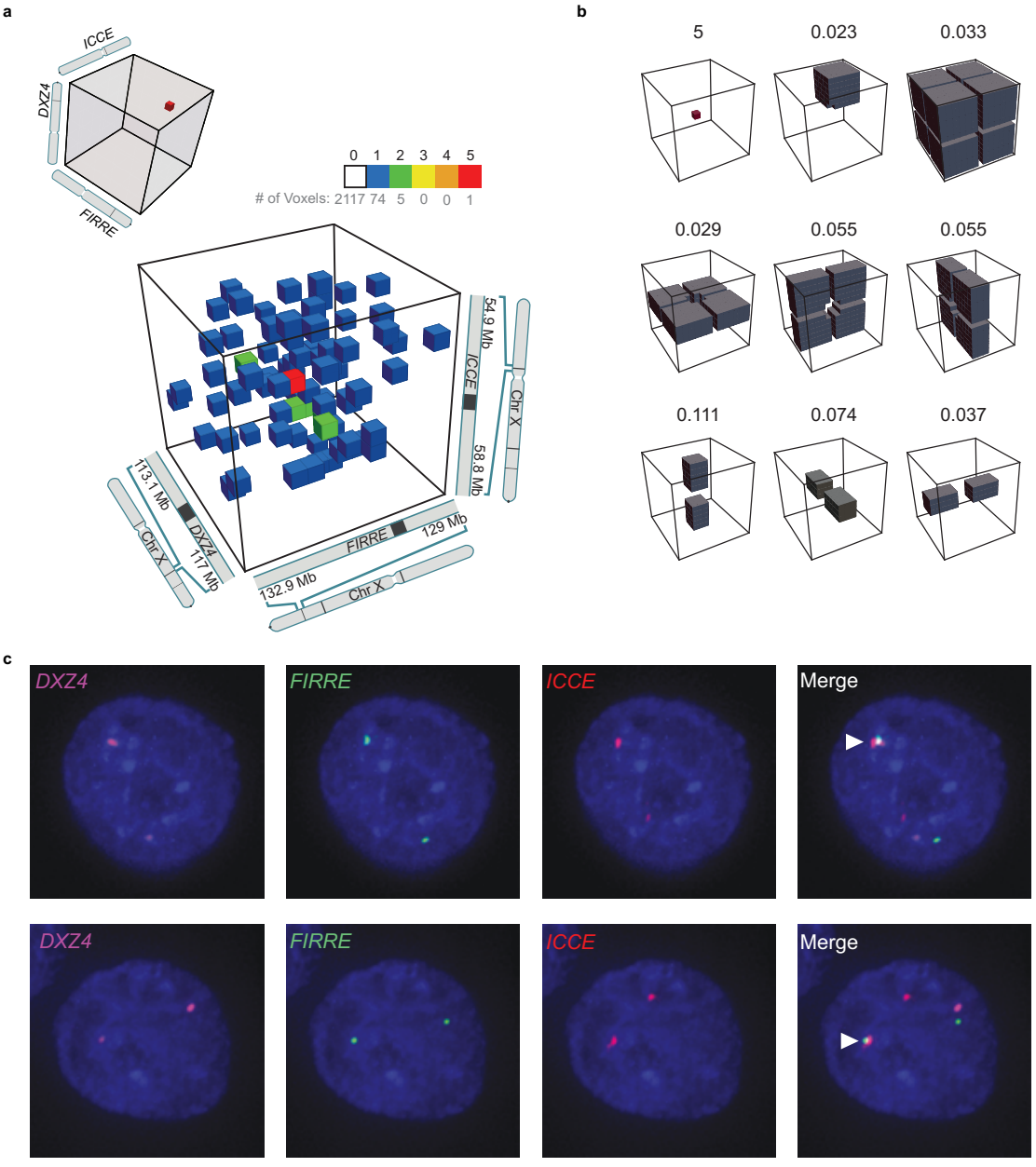
Third, we sought to study the relation between compartment structure and histone modifications on Xi. For this purpose, we chose to study RPE1, a diploid retinal pigment epithelial cell line derived from a human female, because previous studies have characterized the distribution of histone modifications in this cell line. In particular, Xi in RPE1 has been shown to have alternating multi-megabase intervals marked by either histone H3 lysine 9 trimethylation (H3K9me3) or histone H3 lysine 27 trimethylation (H3K27me3); these large intervals are visible as distinct bands during metaphase and the two types of intervals tend to co-localize in the nucleus during interphase[26]. We confirmed these results by performing immunofluorescence using antibodies for H3K9me3 and H3K27me3 both during interphase (Figure 5.4A) and during metaphase (Figure 5.4B). In order to precisely map the genomic intervals decorated by each mark during interphase, we examined RPE1 ChIP-Seq data for H3K27me3 (Figure 5.4B)[106]. We observed a correspondence between the ChIP-Seq signal seen during interphase and the immunofluorescence pattern seen during metaphase, supporting the suggestion that the extent of each mark does not change significantly between the two states[25].

We then used in situ Hi-C to create a diploid contact map of the X chromosome in RPE1. As in our previous high-resolution in situ Hi-C maps from human females, the Hi-C map exhibited superdomains and superloops on Xi but not Xa. Like our previous diploid human map, and unlike our diploid murine map, the RPE1 Xi also exhibited extensive compartmentalization. Compartmentalization patterns vary among cell lines, but we observed that RPE1 had several long compartment intervals on Xi that were not present on Xa, which notably included a contiguous 15 Mb-long stretch from a locus at 100 Mb (denoted x100) all the way to DXZ4 (Figure 5.4B). Strikingly, the compartment

intervals on Xi, reflected in the first principal component of the observed/expected matrix, correspond closely to the alternating intervals of H3K27me3 and H3K9me3 on Xi seen by ChIP-Seq and by immunofluorescence (Figure 5.4B). For example, the 15 Mb region between x100 and DXZ4 corresponded to a contiguous H3K27me3 interval in all 151 Xi chromosomes examined (see Figure 5.4B)[25].

Figure 5.3 *(following page)*: The ICCE-DXZ4 and DXZ4-FIRRE superloops tend to occur simultaneously, forming a hub on Xi. (A) Examination of a small region from the 3D contact tensor of the X chromosome, centered on ICCE, DXZ4, and FIRRE, reveals a peak relative to the local neighborhood. Five contacts are seen in the $(300\text{ kb})^3$ voxel (i.e., 3D pixel) associated with simultaneous co-location of all three loci. There are over 2000 other $(300\text{ kb})^3$ voxels in the region shown; the number of contacts in each is indicated by the color. Of these voxels, five contain 2 contacts each, seventy-four contain 1 contact each, and over 2000 voxels contain no contacts. Note that due to fixed bin width, the voxel size presented in this figure, $(300\text{ kb})^3$, is slightly larger than (and has slightly more contacts than) the exact volume defined ICCE-DXZ4-FIRRE boundaries, which was used for the analyses in the main text. (A) The average frequency of contact in various local neighborhoods surrounding the ICCE-DXZ4-FIRRE peak. The peak is strongly enriched with respect to every model. (C) Representative examples of direct-labeled 3-color DNA FISH in GM12878, a female cell line, showing collocation of ICCE-DXZ4-FIRRE. FISH signals overlay DAPI (blue) and are merged in the far-right panel. White arrowhead indicates 3-way overlap on one X-chromosome. Microscopy did not reveal a single instance of ICCE-DXZ4-FIRRE collocation in male cells.

Figure 5.3: (continued)



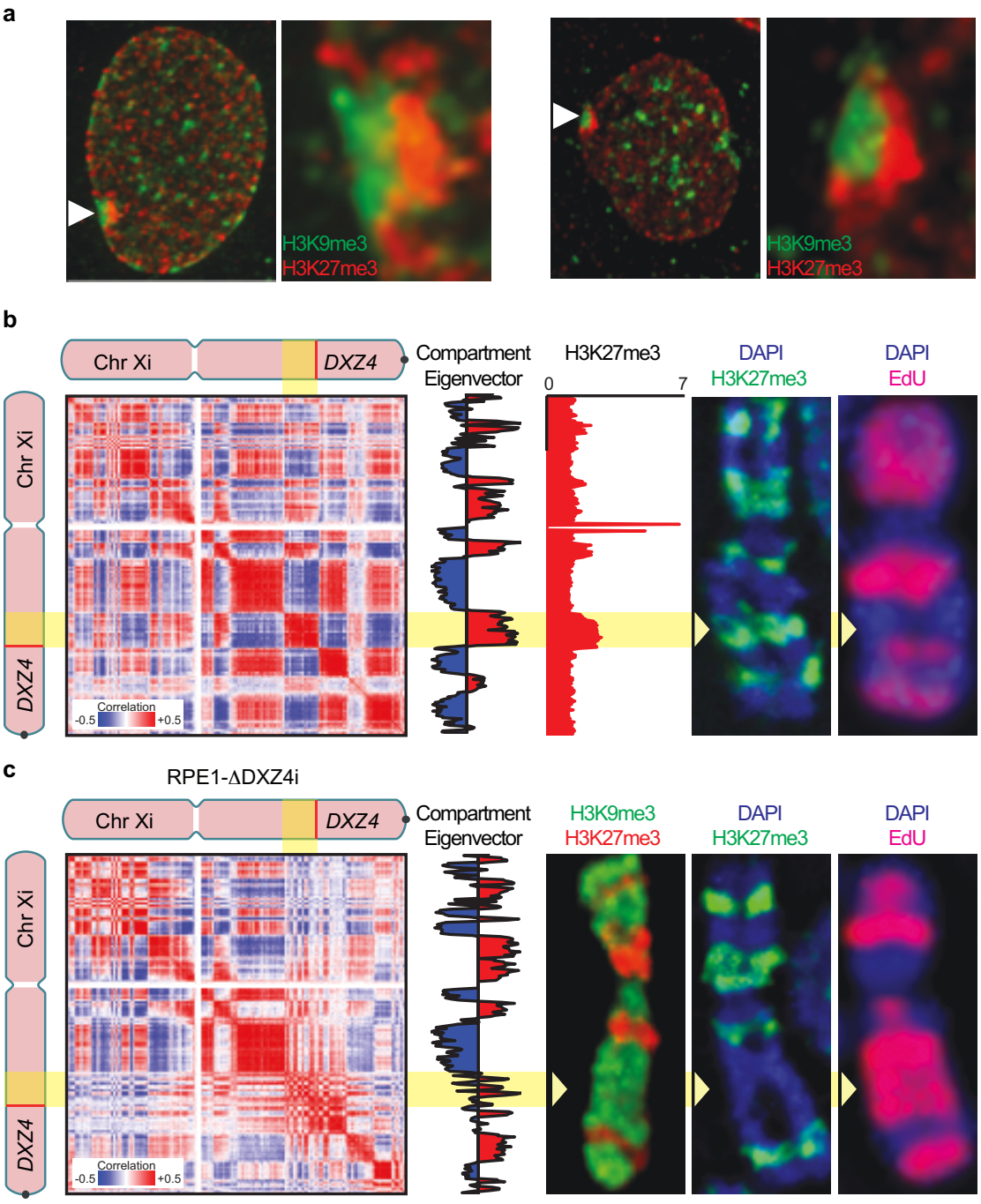
Fourth, we explored the effect of deleting DXZ4 on the structures characterized above. We designed a pair of TALENs (a class of nucleases with a customizable DNA-binding motif) that target two unique inverted repeats flanking the DXZ4 locus. Simultaneous cutting at both ends of the locus resulted in loss of the intervening DNA, including DXZ4. We isolated clones of RPE1 that lacked DXZ4 on either Xi (RPE1-ΔDXZ4i) or Xa (RPE1-ΔDXZ4a). We defined the precise extent of the deletions by sequencing the deletion product. (We created additional RPE1-ΔDXZ4i clones using a dual-guide, CRISPR/Cas9-based genome editing strategy.)

We then used in situ Hi-C to create diploid contact maps of both mutant cell lines (RPE1-ΔDXZ4i and RPE1-ΔDXZ4a), and compared these maps to the maps of wild-type RPE1 generated above (Figure 5.5A,B). RPE1-ΔDXZ4a mutants exhibited the same superdomains and superloops on Xi seen in wild-type RPE1 cells. In contrast, Xi in RPE1-ΔDXZ4i mutants lacked both the superdomains and the superloops anchored at DXZ4 (Figure 5.5B,C). Notably, RPE1-ΔDXZ4i continued to exhibit superloops not anchored at DXZ4, such as the FIRRE-ICCE superloop, suggesting that superloops form independently of one another; we confirmed this observation by using FISH to probe the FIRRE-ICCE superloop (anchor overlap frequency in RPE1-WT: 7 of 114 Xi chromosomes [6%]; in RPE1-ΔDXZ4a: 7 of 107 [7%], and in RPE1-ΔDXZ4i: 2 of 102 [2%]).

When we examined the large contiguous compartment interval between x100 and DXZ4 that is present on Xi in wild-type RPE1 cells, we found that deletion of DXZ4 on Xi (but not on Xa) disrupted the interval's compartmentalization pattern, histone modifications and replication timing. The single compartment interval was replaced by small alternating compartment intervals (Figure 5.4C). Consistent with the observation 6,12 that changes in compartmentalization are associated with changes in histone

Figure 5.4 *(following page)*: Deletion of DXZ4 disrupts compartmentalization, distribution of histone marks, and replication timing. (A) Immunofluorescence showing the distribution of H3K9me3 (green, indirect immunofluorescence) and H3K27me3 (red, direct immunofluorescence) in wild type RPE1 at interphase. White arrowheads indicate the location of Xi that is expanded in the panels to the right, showing the corresponding nuclei. (A) A correlation matrix derived from the RPE1-WT Xi contact map reveals two distinct long-range contact patterns, indicating two subcompartments (1st column). These patterns are reflected in the principal eigenvector, which is shown to the right of the matrix (2nd column). The color of the eigenvector indicate its sign, and thus the long-range pattern exhibited by the corresponding locus (Resolution: 500 kb.) One of these subcompartments correlates well with H3K27me3 ChIP-Seq in RPE1 (3rd column), as well as with a representative metaphase Xi showing the distribution of H3K27me3 (green, indirect immunofluorescence) merged with DAPI (blue) in RPE1 (4th column), and with a representative metaphase Xi showing the pattern of EdU incorporation (red) merged with DAPI (blue) (5th column). The yellow arrowheads indicate the H3K27me3 band between x100 and DXZ4 that replicates earlier in S-phase. (C) A correlation matrix derived from the RPE1-ΔDXZ4i Xi contact map (1st column); its principal eigenvector (2nd column); a representative RPE1-ΔDXZ4i metaphase Xi showing the distribution of H3K9me3 (green, indirect immunofluorescence) and H3K27me3 (red, direct immunofluorescence) (3rd column), a representative metaphase Xi showing the distribution of H3K27me3 (green, indirect immunofluorescence) merged with DAPI (blue) (4th column), and a representative metaphase Xi showing the pattern of EdU incorporation (red) merged with DAPI (blue) (5th column). The compartment interval between x100 and DXZ4 is compromised, and corresponding changes are seen in metaphase histone mark distribution and in replication timing.

Figure 5.4: (continued)



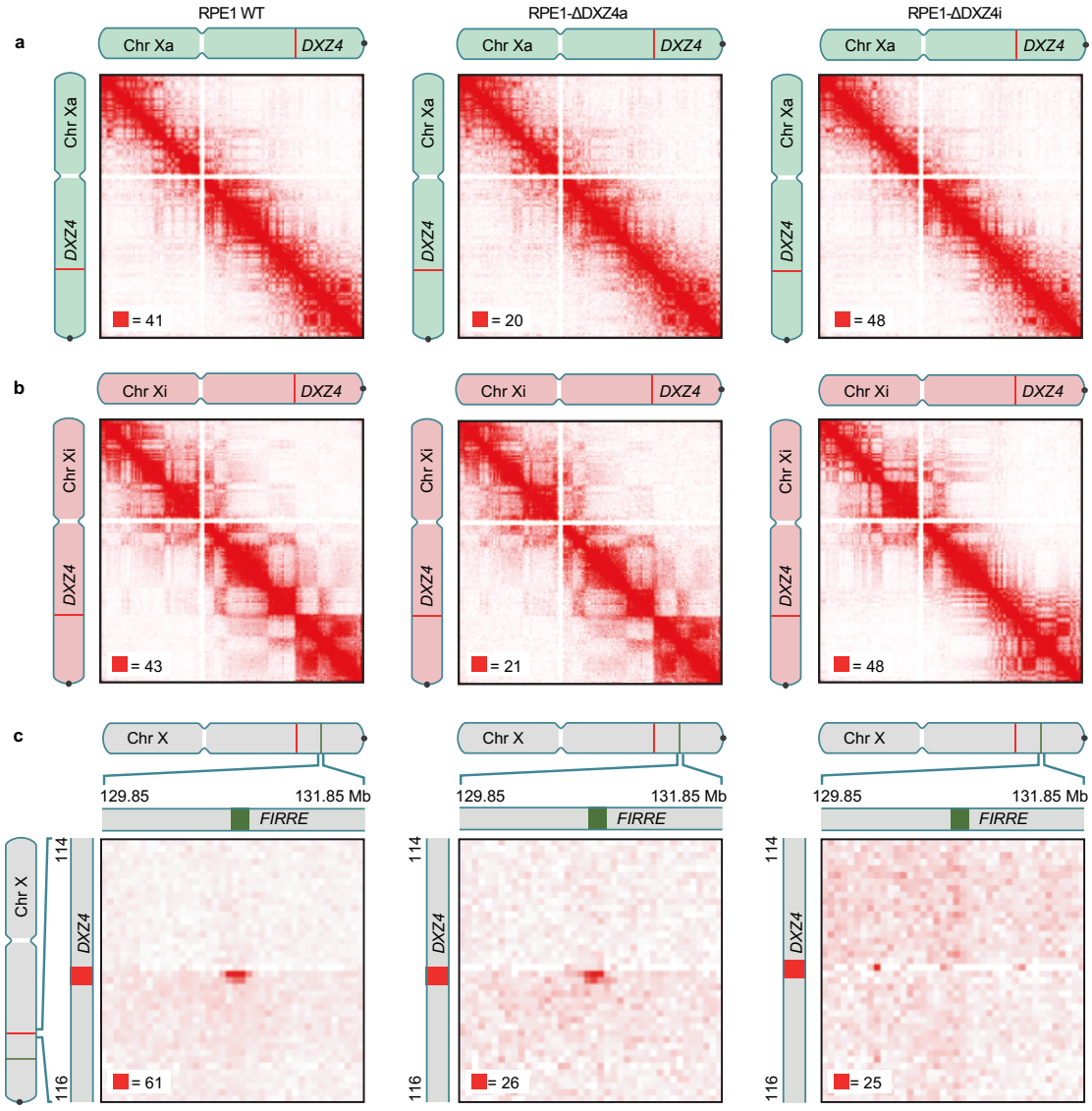


Figure 5.5: Deletion of DXZ4 eliminates Xi superstructure. (A) Maps of Xa in RPE1 (left), RPE1-ΔDXZ4a (middle), and RPE1-ΔDXZ4i (right). Compartmentalization is seen. Superstructure is absent. (A) Maps of Xi in RPE1 (left), RPE1-ΔDXZ4a (middle), and RPE1-ΔDXZ4i (right) exhibit compartmentalization with unusually long compartment intervals. Superdomains are present in wild type RPE1 cells and remain after DXZ4 is deleted on Xa, but not after DXZ4 is deleted on Xi. (Resolution: 500 kb.) (C) The DXZ4-FIRRE superloop is present in wild type RPE1 (left) and in RPE1-ΔDXZ4a (middle), but disappears in RPE1-ΔDXZ4i (right). (Resolution: 50 kb.)

modification patterns, the contiguous interval of H3K27me3 across this interval was either significantly diminished or completely absent in three independent RPE1-ΔDXZ4i clones (Clone 1 Frequency: 29.1%, n=148; Clone 2 Frequency: 23.8%, n=172; Clone 3 Frequency: 1.0%, n=97; p = <0.0001 Fisher's exact test; Figure 5.4B) and the interval was instead decorated by H3K9me3 (Figure 5.4B). By using 5-Ethynyl-2'-deoxyuridine (EdU) labeling, we found that DNA in this interval also replicates much later in RPE1-ΔDXZ4i cells than in the wild-type RPE1 cells (Figure 5.4A,B). In contrast to these findings, we observed no major effect on overall transcription of Xi. RNA-Seq showed expression changes exceeding 2-fold in only 10 of 982 genes, after FDR correction. The altered genes had no obvious significance. Similarly, RNA-FISH of 12 genes on Xi revealed no consistent changes across RPE1-ΔDXZ4i clones.

The results above show that DXZ4 plays an important role in the nuclear architecture of Xi. A key question is how DXZ4 brings about the presence of superloops and superdomains on Xi. The DXZ4 locus typically comprises between 10 and 100 copies of a repeat element containing a CTCF motif[24]. These CTCF motifs are the only sequences in DXZ4 showing evidence of strong purifying selection[63]. DXZ4 on Xi is known to be hypomethylated and bind CTCF, while on Xa it is hypermethylated and does not bind CTCF[24].

Recently, we proposed a model in which CTCF sites mediate the formation of chromatin loops and domains via extrusion[123]; see also[3,47,102]). In this model, an extrusion complex comprising two DNA-binding subunits is loaded onto chromatin at a single genomic position. The subunits slide along chromatin in opposite directions, thereby extruding a loop, and stop when they arrive at a CTCF site in the appropriate orientation, which serves as an 'anchor' (Figure 5.6A). Simulations showed that extruded loops naturally give rise to domains and that knowledge of the CTCF-bound sites is sufficient

to predict, with reasonable accuracy, the results of Hi-C experiments.

Although the model above assumed extrusion complexes stop completely at appropriately oriented CTCF sites, similar results are obtained if the complexes undergo partial arrest for a period of time. If so, the model would also account for the presence of superloops and superdomains at loci on Xi that contain a large number of CTCF-binding sites: extrusion complexes would be sequestered at such loci for a much longer period of time than at typical CTCF sites. In contrast, these structures would not be seen on Xa, where DXZ4 is packaged into constitutive heterochromatin[24] and does not bind CTCF and therefore could not arrest the extrusion complex.

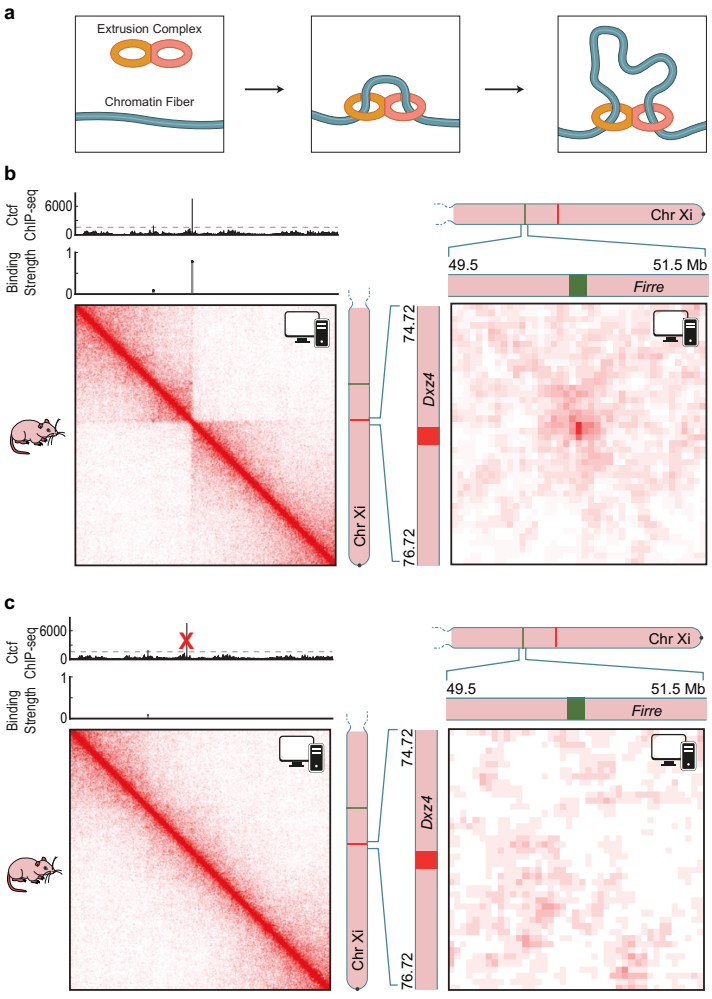
To test this hypothesis, we modeled the mouse Xi as a long homopolymer containing CTCF binding sites, whose positions were derived from Patski Ctcf ChIP-Seq data[14]. We used molecular dynamics simulations to examine the behavior of this polymer in a solvent containing extrusion complexes, and used the resulting ensemble to generate a contact map for Xi (Figure 5.6B). We found that the contact map was partitioned into two superdomains, whose boundary lay at Dxz4. It also clearly exhibited the Dxz4-Firre superloop. When we simulated the effects of deleting Dxz4, these features disappeared (Figure 5.6C). Thus, Hi-C data for Xi is consistent with a model in which DXZ4 brings about the formation of superdomains and superloops by serving as a tandem anchor during loop extrusion.

5.3 Conclusion

In summary, we find that the position of DXZ4 with respect to superloop anchors and the superdomain boundary on Xi is broadly conserved across eutherian mammals; that DXZ4, FIRRE, ICCE, and possibly other loci simultaneously co-locate to form a superhub comprising superloop anchors on Xi; and that the deletion of DXZ4 broadly disrupts the chromatin structure of Xi by eliminating superdomains, disrupting superloops anchored at DXZ4, and modifying the distribution of chromatin marks and compartments. Finally, we show that the behavior of DXZ4, which contains numerous CTCF binding sites, resembles the behavior of individual CTCF sites, but produces features — loops, domains, histone mark patterns, and compartmentalization — at the whole-chromosome scale. Despite the larger feature size, the observations appear to be consistent with extrusion as an underlying mechanism, suggesting that genomes may exhibit similar folding principles across a wide range of scales.

Figure 5.6 *(following page)*: Extrusion Model of Superstructure. (A) In our physical model of loop formation by extrusion[123], an extrusion complex comprising two DNA-binding subunits is loaded onto chromatin. The subunits form a loop by sliding in opposite directions. When they arrive at an anchor site, they have a probability of binding and thus halting the extrusion process. (A) We generated an ensemble of Patski Xi chromosome configurations using the extrusion model, with anchor binding probabilities drawn from Ctcf ChIP-Seq data in Patski. We calculate a contact map from this ensemble. A superdomain boundary is formed at Dxz4 (left); a superloop forms between Dxz4 and Firre (right). (C) Simulating the deletion of Dxz4 leads to disappearance of the superdomain (left) and superloop (right).

Figure 5.6: (continued)



6

The Information Capacity of Specific Interactions

6.1 Introduction

Specific interactions between many species of components is the bedrock of biochemical function, allowing signal transduction along complex parallel pathways and self-assembly of multi-component molecular machines. Inspired by their role in biology, engineered specific interactions have opened up tremendous opportunities in materials synthesis, achieving new morphologies of self-assembled structures with varied and designed functionality. The two major design approaches for programming specific in-

teractions either use chemical specificity or shape complementarity.

Chemical specificity is achieved by dividing binding sites into smaller regions, each of which can be given one of A "colors", or unique chemical identities. Sites bind to each other based on the sum of the interactions between corresponding regions. For example, a recent two color system paints the flat surfaces of 3-dimensional polyhedra with hydrophobic and hydrophilic patterns[118], or with a pattern of solder dots[55], allowing polyhedra to stick to each other based on the registry between their surface patterns. Another popular approach uses DNA hybridization, where specific matching of complementary sequences has been used to self-assemble structures purely from DNA strands[145,74], and from nanoparticles coated with carefully chosen DNA strands[95,4,141,15,107].

Shape complementarity uses the shapes of the component surfaces to achieve specific binding, even though the adhesion is via a nonspecific, typically short-range potential. In the synthetic context, shape-based modulation of attractive forces over a large dynamic range was first proposed and experimentally demonstrated for colloidal particles[88,89] using tunable depletion forces[156,157]. Recent experiments have explored the range of possibilities opened up by such ideas, from lithographically designed planar particles[61] with undulating profile patterns to "Pac-man" particles with cavities that exactly match smaller complementary particles[122]. The number of possible shapes that can be made using these types of methods depends on fabrication constraints but the possibilities can be quite rich[151,40]. Using only non-specific surface attraction, experiments have achieved numerous and complex morphologies such as clusters, crystals, glasses, and superlattices[155,120,88,96,150].

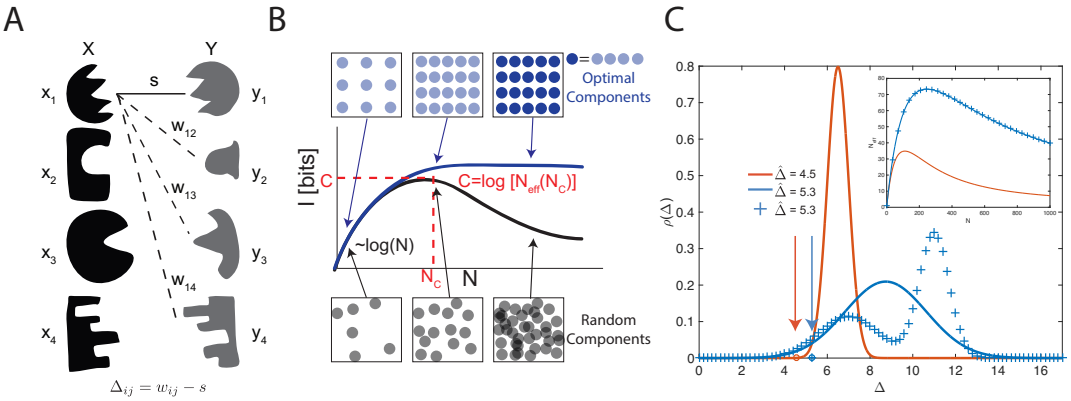
A further class of programmable specific interactions combines both chemical specificity and shape complementarity. The canonical example is protein binding interactions[70]; the binding interactions between two cognate proteins are specified by their

amino acid sequence, which programs binding pockets with complex shape and chemical specificity. Recent efforts[75,80] aim to rationally design these protein interactions for self-assembly. Since both the shape of the binding pocket and its chemical specificity is determined by the same amino acid sequence, these two features cannot be controlled independently. Other synthetic systems offer the promise of independent control of chemical and shape binding specificity, giving a larger set of possible interactions.

These diverse systems achieve specific interactions through disparate physical mechanisms, with different control parameters for tuning binding specificity. However, they must all solve a common problem[113,132]: create a family of N "lock" and "key" pairs that bind well within pairs but avoid off-target binding across pairs ('crosstalk'). Any crosstalk limits the efficacy of the locks and keys. For example, in the context of DNA-based affinities, although there are 4^L unique sequences of length L, the strong off-target binding severely restricts the number that can be productively used. Analogously, for colloidal systems driven by depletion interactions, there can be significant off-target binding due to partial contact. The performance of a system of specific interactions depends acutely on how the system constraints (e.g. number of available bases, fabrication length scale, etc.) limit its ability to avoid crosstalk.

In this paper, we develop a general information theory-based framework for quantitatively analyzing specificity in both natural and synthetic systems. We use a metric based on mutual information to derive a bound on the number of different interacting particles that a system can support before crosstalk overwhelms interaction specificity. Increasing the number of nominally distinct pairs beyond this limit cannot increase the effective number of distinguishable species. We compute this information-theoretic 'capacity' for different experimental systems of recent interest, including DNA-based affinities and colloidal experiments in shape complementary. We show that shape-based coding fun-

Figure 6.1: Information theory determines the capacity of systems of specific interactions. **A)** A model system of locks (black) which each bind with energy s to their specific key (gray) via some specific interaction. **B)** As the number N of lock-key pairs is increased, non-cognate locks and keys inevitably start resembling each other as they fill up the finite space of all possible components (square boxes), both with optimized or random design of lock-key pairs. Consequently, mutual information I between bound locks and keys rises with N for small N but reaches a point of diminishing returns at $N = N_C$; due to the rapid rise in off-target binding energy, I can no longer increase, and for randomly chosen pairs, will typically decrease. The largest achievable value of I is the 'capacity' C. **C)** Capacity C can be estimated from the distribution $\rho(\Delta)$ of the gap $\Delta = w - s$ between off-target w and on-target binding energy s for randomly generated lock-key pairs. Among the three distinct $\rho(\Delta)$ shown, the blue distributions have the same $\beta\hat{\Delta} = -\log\langle\beta\Delta e^{-\beta\Delta}\rangle_\rho$. Inset: $I(N)$ (c.f., Eqn. 6.3), is the same for the blue distributions which, despite being markedly different in shape, have the same $\hat{\Delta}$, which captures the essential aspects of crosstalk.



damentally results in lower crosstalk and higher capacity than color-based coding. We also find that shape and color-based coding can be combined synergistically, giving a super-additive capacity that is greater than the sum of the color and shape parts.

6.2 The Capacity of Random Ensembles

We consider systems where every component is designed to interact specifically with a single cognate partner, while interactions between "off-target" components are undesirable crosstalk. We assume that N distinct "locks" x_1, x_2,..., $x_N \in X$, have unique binding partners, "keys" y_1, y_2,..., $y_N \in Y$ (Fig. 6.1A). The physics of a particular

system determines the binding energy $E_{ij} \equiv E(x_i, y_j)$ between every lock and key. Assuming equal concentrations of locks and keys in a well-mixed solution, binding between lock x_i and key y_j will occur with probability $p(x_i, y_j) = e^{-\beta E_{ij}}/Z$ where Z is a normalization factor such that $\sum_{i,j} p(x_i, y_j) = 1$ (see Appendix A) and $\beta^{-1} = k_B T$ is the temperature scale. The mutual information $I(X;Y)$ transmitted through binding is defined as

$$I(X;Y) = \sum_{x_i \in X, y_j \in Y} p(x_i, y_j) \log_2 \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \tag{6.1}$$

where $p(x_i)$ is the marginal distribution of x_i, representing the total probability of seeing x_i in a bound pair (and similarly $p(y_j)$). Mutual information $I(X;Y)$ is a global measure of interaction specificity in systems with many distinct species; it quantifies how predictive the identity of a lock x_i is of the identity of a key y_j found bound to it.

Consider a set of interacting lock-key pairs for which $E_{ii} = s$ for all cognate pairs (strong binding), while for crosstalking interactions (weak binding) $E_{ij} = w_{ij} = s + \Delta_{ij}$. We assume Δ_{ij} are i.i.d. random numbers drawn from a distribution of gap energies $\rho(\Delta)$, with $\Delta > 0$, where the exact form of $\rho(\Delta)$ depends on the physics of the system. Denoting $\langle\rangle$ as an average with respect to $\rho(\Delta)$, one can approximate Eqn. A.8 as

$$I = \log_2 N_{eff}(N), \tag{6.2}$$

$$N_{eff}(N) = \frac{N}{1 + (N-1)\langle e^{-\beta\Delta}\rangle} e^{-\frac{(N-1)\langle \beta\Delta e^{-\beta\Delta}\rangle}{1+(N-1)\langle e^{-\beta\Delta}\rangle}} \tag{6.3}$$

(see Appendix A). In a system with crosstalk that contains N nominally distinct lock-key pairs, $N_{eff}(N)$ is the effective number of fully distinguishable lock-key pairs. N_{eff} can be much smaller than N if crosstalk is significant (e.g., if $\langle e^{-\beta\Delta}\rangle \sim O(1)$).

Intuitively, information theory predicts that a system with $N_{eff}(N)$ non-crosstalking

lock-key pairs can perform a task with the same effectiveness as a system with N crosstalking species. For example, in the self-assembly of a multi-component structure, distinct but crosstalking species can take each other's place, decreasing the effective number of species. This effect has been shown to reduce self-assembly yield[98,68,59]. Similarly, the efficacy of N parallel signaling pathways is known to be reduced by crosstalk[116]. In Fig. 6.1B we show a typical plot of $I = \log_2 N_{eff}(N)$. N_{eff} grows initially with N, but stops growing at $N \sim N_C$, the point of diminishing returns; adding any further species beyond N_C only increases the superficial diversity of species but cannot increase N_{eff}.

Paralleling Shannon's theory of communication, we define 'capacity' C as

$$C \equiv \max_N I = \log_2 N_{eff}(N_C) \tag{6.4}$$

(Fig. 6.1B)*. The capacity is the largest number of bits of information that can be encoded using a system of specific interactions and still be uniquely resolved by the physics of interactions. Determining C, or equivalently the largest value of N_{eff}, is of crucial importance to both synthetic and biological systems since it limits, for example, the number of independent signaling pathways or the complexity of self-assembled structures.

We can compute capacity for any crosstalk energy distribution $\rho(\Delta)$ by finding the maximum of Eqn. 6.3. A useful approximation is

$$C \approx -\log_2\langle \beta\Delta e^{-\beta\Delta+1}\rangle, \tag{6.5}$$

*'Capacity' in information theory is often measured in bits/second, whereas here we intentionally use the same units as I. Furthermore, capacity is traditionally defined as a maximum over all possible distributions $p(X)$; here we restrict to maximizing only over one parameter, N, where all N pairs are randomly chosen from the ensemble (see Appendix).

giving a simple rule for the dependence of capacity on the binding energy distribution (see Appendix A). The importance of maximizing $\beta\hat{\Delta} \equiv -\log\langle\beta\Delta e^{-\beta\Delta}\rangle$ in Eqn. 6.5 is intuitive: in order to increase the capacity of the system, the (exponential average of the) gap between on-target $\beta\bar{s} \equiv -\log\langle e^{-\beta s}\rangle$ and off-target binding $\beta\bar{w} \equiv -\log\langle e^{-\beta w}\rangle$ should be made as large as possible (see Appendix A for precise relationship between $\hat{\Delta}, \bar{s}$ and $\bar{w}$). Fig. 6.1C shows three distinct probability distributions, two of which have identical $\hat{\Delta}$. As predicted, N_{eff} reaches a higher maximum for distributions with larger $\hat{\Delta}$.

We note that our definition of capacity uses equilibrium binding probabilities and hence applies only at long times compared to unbinding times. In practice, this typically limits $|s| \leq 10\ k_B T$, and so we use this bound on s herein. The formalism can be easily extended to include kinetic effects by computing $p(x_i, y_j)$ at a finite time t, though this is not our focus here.

In what follows, we show how capacity depends on binding interactions and fabrication constraints for several systems of recent interest. In most systems, the on-target binding energy typically strengthens with the binding surface area S of cognate pairs as $\bar{s} = -\epsilon S$, where ϵ is the binding energy per unit area. However, we find that the off-target energies $\bar{w}$ can grow with S at very different rates across several systems we study. We parameterize this variation as

$$\bar{w} = -\epsilon\alpha S^{\gamma} \tag{6.6}$$

where α, γ depend on the details of binding interactions. We show below that if the specificity is determined purely by 'colors' (i.e., chemical identities), then $\gamma = 1$. In contrast, if specificity arises from shape complementarity, $\gamma \approx 0$, as long as the range

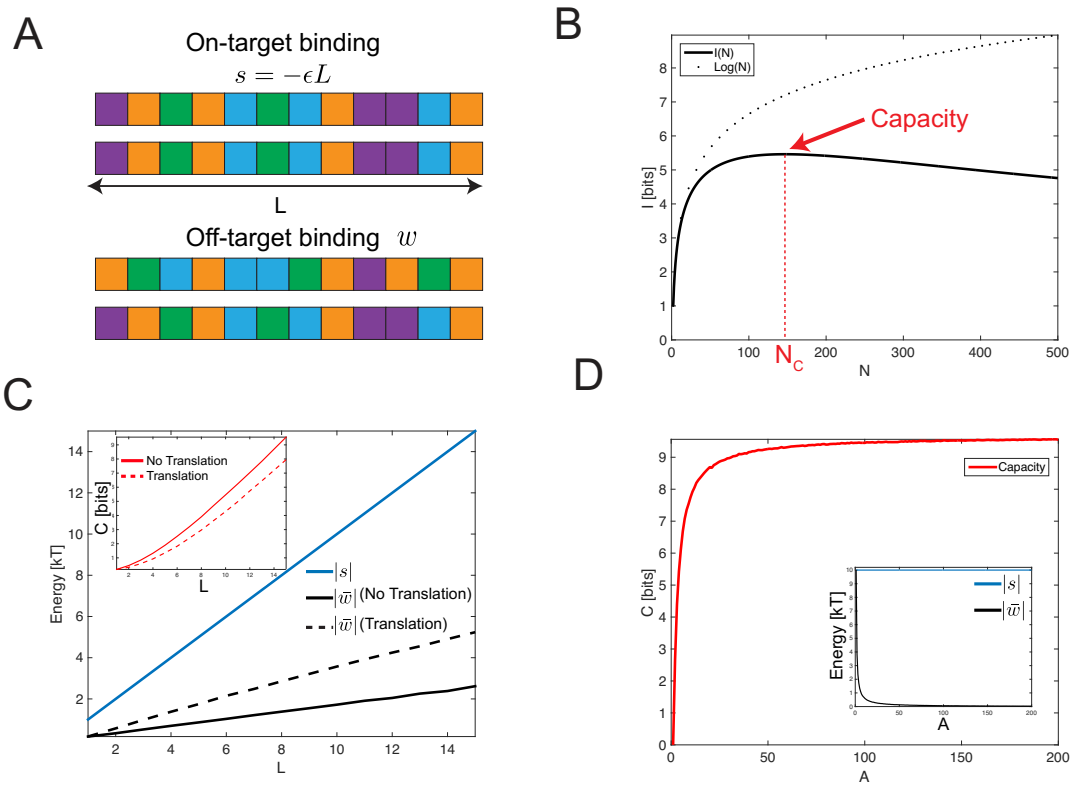
of the surface attraction is small compared to the length scale of shape variation. Thus crosstalk grows very slowly with the number of independent binding units in shape-based systems, allowing for a dramatic decrease in crosstalk and improvement of capacity relative to systems that use chemical specificity.

6.3 The Capacity of Color

We first consider the capacity of interactions mediated through binding sites which are subdivided into multiple regions, each of which can be assigned any one of A chemical identities or "colors". We take inspiration from DNA coding that acts via complementary hybridization between single stranded DNA. Previous work[93] developed engineering principles for determining the optimal length and nucleotide composition of these DNA strands based on detailed models of the binding energy. Information theoretic measures have also been used to understand binding of transcription factors to DNA and other sequence-based molecular recognition problems[99,129,67,125,147]. While the theory of DNA coding has a long history[1], our contribution here formulates the problem in a mutual information framework that relates the capacity to a physical quantity and hence allows for direct comparison of varied chemical ('color') and shape systems.

In our simplified color model, a lock is composed of L units, each of which is painted with one of A chemical colors (Fig. 6.2A). Each color binds to itself with energy $-\epsilon$ and binds to other colors with energy 0, such that locks and their cognate keys have the same sequence. The binding energy of any two strands x_i and y_j is given by $E_{ij} = \sum_{l=1}^{L} -\epsilon\delta_{x_i^l, y_j^l}$ (where x_i^l is the color of the lth site of x_i, and δ is the Kronecker delta). We analyze this system with translations, where E_{ij} is given by the strongest binding across all possible translations of the two strands relative to each other, as well as without translations.

Figure 6.2: Complementary color components demonstrate the capacity of programmable interactions. **A)** Each lock x_i has L distinct units, each of which can be one of $A = 4$ 'colors' or chemical identities. Each color has a strong affinity ϵ for itself. Cognate locks and keys have the same sequence of colors and bind with energy $s = -\epsilon L$. Off-target binding energy w given by the number of accidental color matches; $w = -5\epsilon$ in the example shown. **B)** Mutual information as a function of N, the number of lock-key pairs. I increases initially as $\log(N)$, but then reaches a maximum value, the capacity, and then decreases. ($L = 10$, $\epsilon = 1\ k_BT$, $A = 4$). **C)** Increasing L increases both the on-target strength $|s|$ and the off-target strength $|w|$. Inset: Capacity scales linearly with L. If translation is allowed, $|s|$ is unaffected, but $|w|$ is higher (black, dashed), and therefore the capacity is lower (red, dashed). ($L = 10$, $\epsilon = 1\ k_BT$, $A = 4$). **D)** Increasing the alphabet size A does not affect on-target binding s, but does decrease $|w|$, thereby increasing $\hat{\Delta}$. ($L = 10$, $\epsilon = 1\ k_BT$)



We calculate $I(N)$ by sampling N randomly selected pairs of locks and keys, constructing the interaction matrix E, and computing $I(N)$ using Eqn. A.8. We average $I(N)$ over many repetitions. An approximate but faster method to compute $I(N)$ (necessary for large L, N) uses Eqn. 6.3, sampling random pairs of off-target locks and keys to estimate $\langle e^{-\beta\Delta}\rangle$ and $\langle \beta\Delta e^{-\beta\Delta}\rangle$. The two methods give nearly identical results (see Appendix A), and the calculations in the paper henceforth are carried out with the second method.

Fig. 6.2B examines $I(N)$ when $L = 10$, $A = 4$, and $\epsilon = 1\ k_BT$ so that a lock and its key bind together with on-target energy $\beta s = -10$. The mutual information has a maximum of 5.5 bits near $N_C = 146$, far less than the total number of unique sequences ($4^{10} = 1,048,576$). Due to crosstalk, even though there are nominally 146 pairs at capacity, the system behaves as if there are only $N_{eff} = 44$ independent pairs.

An obvious way of increasing capacity is to boost $\hat{\Delta}$ by increasing L. This strengthens both on-target binding and off-target binding, since both s and w scale with L ($\beta\bar{w} = -L\log\frac{A-1+e^{\beta\epsilon}}{A}$). However, the gap between them widens, and the capacity scales linearly with L (Fig. 6.2C solid line) (see Appendix A). As a comparison, we also show the capacity when translation is allowed between any two strands. Off-target strands can now translate until they find the strongest binding, increasing crosstalk and thus lowering capacity.

In practice, on-target binding $|s|$ must be limited to below approximately 10 k_BT for the binding to be reversible; hence L cannot be increased arbitrarily without also decreasing ϵ. An alternate way to increase capacity at fixed s is to increase the number of 'colors' A. As $A \to \infty$, accidental mismatches in off-target binding are rare; $|w| \to 0$, and the capacity is only limited by s. In Fig. 6.2D, capacity in the large A limit can be approximated by setting $\beta\Delta = -s = 10$ in Eqn. 6.5, giving $C = 9.6$ bits.

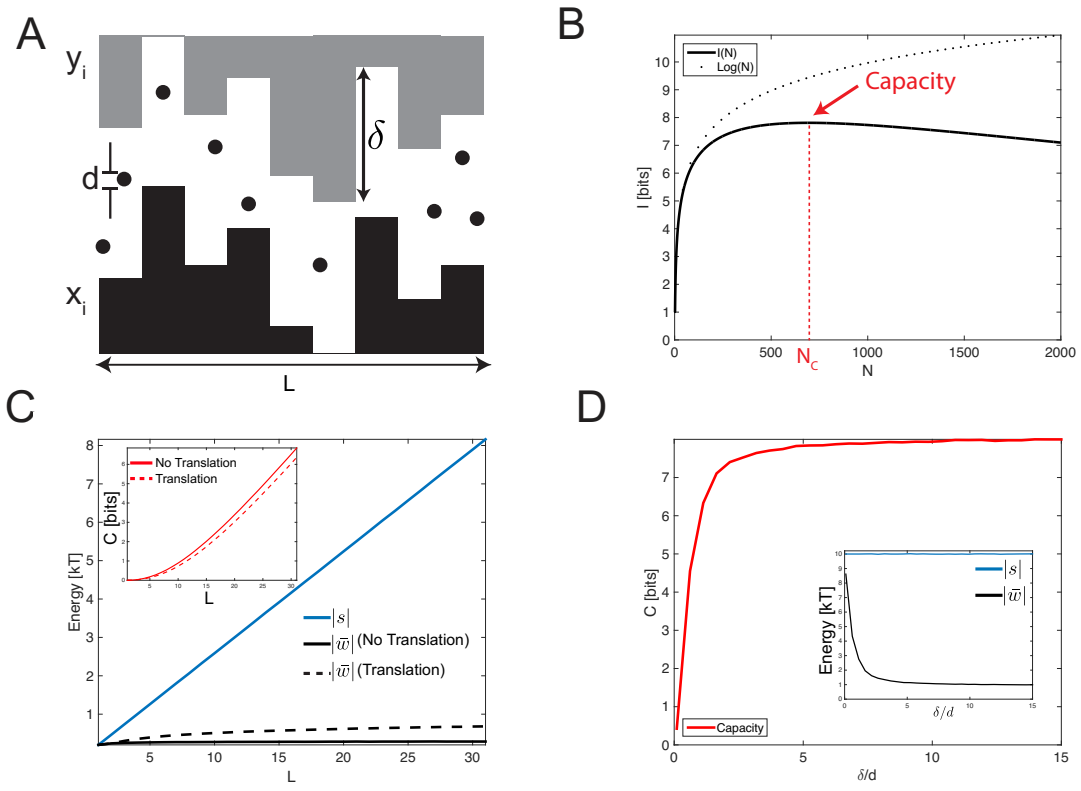
However, in practice, alphabet size A cannot be easily increased in experiments, and other techniques must be used to decrease the off-target binding strength, such as the use of shape complementarity.

6.4 The Capacity of Shape

Systems of interacting, complementary shapes are characterized by the non-specific binding of surfaces mediated by a short-range force of characteristic length λ_{shape}. The components' shapes sterically allow or inhibit two surfaces from coming into contact, dictating specificity. We find that crosstalk is qualitatively weaker in such shape-based systems, resulting in higher capacity than color-based models.

We examine the capacity of a model inspired by a recent experimental system consisting of lithographically sculpted micron-sized particles with complementary shapes[61] whose attractive interactions are mediated by the depletion force. The constraints on the shapes of these components (size < 10 μm, line width > 400 nm, radius of curvature > 200 nm) still leave a large variety of shapes that can interact in a lock and key fashion, yet crosstalk between similarly shaped components reduces the number of effectively unique pairs. We model this system by defining each solo component as a series of L adjoining bars of various heights, whose profile is similar to a Tetris piece. For each lock x_i, the shape of the cognate key y_i is exactly complementary, as in Fig. 6.3A. We account for fabrication constraints by setting the width of each bar to 1 μm and restricting the change of one bar height relative to its neighboring bars to be less than $\delta = 1$ μm. Depletant particles of diameter d (typically $100 - 200$ nm) create an attractive energy of $-\epsilon(d-h)$ for two surfaces separated by $h < d$. Thus $\lambda_{\text{shape}} \sim d$. In experiments, ϵ is set by the depletant particle volume fraction and the temperature. In principle, the fabrication fidelity must also be accounted for, as local defects in the

Figure 6.3: Complementary shapes demonstrate the capacity of programmable interactions. **A)** Each shape x_i is made of L vertical bars of different heights and has a corresponding binding partner y_i shaped exactly as its complement; interactions are mediated via depletant particles of diameter d, and each adjoining bar can change by a maximum amount of δ. **B)** Mutual information as a function of N, the number of lock-key pairs, showing a capacity of 7.8 bits. ($L = 10$, $d = 0.2$ μm, $\delta = 1$ μm, $\bar{s} = -10$ k_BT). **C)** Increasing L increases the on-target binding strength $|s|$ (blue), but has little effect on off-target binding strength $|w|$ (black), unlike with colors (Fig. 6.2C). Inset: Capacity scales linearly with L. Allowing translations has no effect on $|s|$, but increases $|w|$ (black, dashed) and therefore decreases C (red, dashed). ($d = 0.07$ μm, $\delta = 1$ μm, $\epsilon = 1$ k_BT). **D)** Fixing $\bar{s} = -10$ k_BT, the capacity can be increased by decreasing δ: when δ/d is small, on-target keys are indistinguishable from off-target keys, and so capacity is small. Increasing δ decreases the crosstalk, and capacity increases accordingly.



shape will disrupt cognate binding. The effect of such defects is shown in Appendix A; we find that defects of size much less than d, the depletant particle size, have minimal impact on capacity. We assume such a limit in the remainder of the text.

We find that crosstalk with shapes differs fundamentally from the color models discussed earlier. While on-target binding strength still increases linearly with L, off-target binding is almost independent of L (Fig. 6.3C). In fact, we find that for large enough L, off-target binding $\bar{w} \sim -L^0$; for larger δ/d (or smaller L), $\bar{w}$ is still strongly sublinear in L (see Appendix A). The weak dependence of $\bar{w}$ on L can be understood intuitively, as a lock pressed to a random mismatched key will typically come into contact at a single location. In contrast, in color-based systems, off-target locks and keys are in full contact and hence $\bar{w} \sim -L$. Thus $\hat{\Delta}$ and hence capacity C for shape systems can be significantly higher than for color based systems with the same strong binding energy $\bar{s}$. In Fig. 6.3B, $C_{shape} = 7.8$ bits while $C_{color} = 5.5$ bits with similar parameters ($\bar{s} = -10\ k_BT$, $L = 10$). Finally, in Fig. 6.3D, for fixed L, we find that capacity falls rapidly and all specificity is lost when the spatial range of depletion interactions $\lambda_{\text{shape}} \sim d$ exceeds the scale of spatial features δ, as expected. These results are consistent with earlier experiments[156] and computational models[157] that established a high dynamic range in the strength of depletion interactions between surfaces roughened by asperities, and in particular found that the attraction between surfaces was diminished when the asperity height was below the depletion particle size.

Our results, while intuitive in retrospect, point to a qualitative advantage for coding through shapes; random mismatched shapes have a crosstalk that is, at worst, sublinear in binding site size while crosstalk is linear in site size for color-based systems. Our work suggests that such increased specificity is very robust as it is derived from basic properties of shape itself. Knowing the precise benefits of shape-based coding is important

in deciding to incorporate it in engineering efforts going forward.

6.4.1 Lock and Key Colloids

We may further apply this framework to the recent experimental system of lock-key colloids. In this system[122], a key is a sphere of radius r (typically 1-3 microns), while its cognate lock is a larger sphere with a hemispherical cavity of radius r, complementary to its key (see Appendix A). The attraction is mediated by depletant particles of diameter $d \approx 50 - 100$ nm. Multiple pairs of locks and keys may be used concurrently, with the ith pair having a key radius of r_i, with the risk of keys binding to incorrect locks.

How should one choose N lock-key radii r_i to minimize crosstalk and maximize capacity? We may gain some intuition by considering a system containing only two lock-key pairs of radius r_1 and r_2 respectively. The on-target binding energies of the two pairs are proportional to the area of contact: $E_{11} \sim r_1^2$, $E_{22} \sim r_2^2$ since each key makes perfect contact with its own lock. Assuming $r_1 < r_2$, crosstalk $E_{12} \sim r_1$, corresponding to the larger key of size r_2 contacting an annulus around the smaller lock of size r_1. The other crosstalk energy $E_{21} \sim r_2^2 - (r_2 - r_1)f(r_2, d)$ is typically much larger, corresponding to the smaller key fitting into the larger lock of size r_2 (see Appendix A for complete derivation). Thus, there are two competing pressures on the radii r_1, r_2: increasing the overall size of both pairs r_1, r_2 improves specificity since the on-target energies r_1^2, r_2^2 grow faster than the crosstalk terms. Yet E_{21} grows rapidly if the radii are too similar to each other. Hence the optimal solution for $N = 2$ requires setting $r_2 = R_{max}$ (the largest allowed radius) and $r_2 - r_1 \approx d$. The binding energy of 6 particles (in this case optimally chosen to maximize I) is shown in Fig 6.4A, with on-target binding and the two types of off-target binding shown.

This intuitive argument does not capture many-body effects that determine capacity

for larger N. We find the optimal $\{r_i\}_N$ at fixed N by maximizing the mutual information I in Eqn. A.8 numerically through gradient descent; note that Eqns. 6.3 cannot be used since the on-target binding energy s varies across pairs. Fig. 6.4B (solid line) shows the mutual information of optimally chosen radii as a function of N, an improvement over randomly chosen radii (dashed line). Fig 6.4C shows the optimal set of radii for various N, with $d = 100$ nm and $R_{max} = 3\ \mu m$; the optimal spacing of the radii is $O(d)$.

Interestingly, when $N > 6$, the system has exceeded its capacity. I does not increase any further (Fig 6.4B) and the optimal set involves repeating locks and keys of the smallest radii. Intuitively, the smallest lock-key pairs have become so small that making an additional lock-key pair of an even smaller radius would yield very low self-binding energy relative to the incurred crosstalk. Hence the only way to increase N without decreasing I is to create new nominal pairs at the smallest radius; such pairs are obviously indistinguishable through physical interactions and hence do not increase mutual information any further. We find that this capacity decreases with increasing size of depletant particles and falls to $N_C \sim 1$ by $d = 400$ nm. Similarly, increasing R_{max} (with fixed largest cognate binding energy $s = -8\ k_BT$) increases capacity.

Thus we find that this colloidal particle system can support about $N_C \sim 6-8$ lock-key pairs without much crosstalk, with depletion particles of diameter 100 nm and restricting the largest binding energy to $s_{\max} = -10\ k_BT$. This is far smaller than the capacity of either DNA sequences or general shape-based strategies. However, these lock-key colloidal pairs are characterized by only one parameter (the radius), so the space of available pairs is significantly smaller than DNA or shape systems with L parameters. In particular, in the current system, additional lock-key pairs are forced to be of smaller radii and hence of lower and lower cognate binding energies. Such considerations emphasize the importance of quantitative information-theoretic optimization in systems

with such a limited shape space.

6.5 Combining Channels

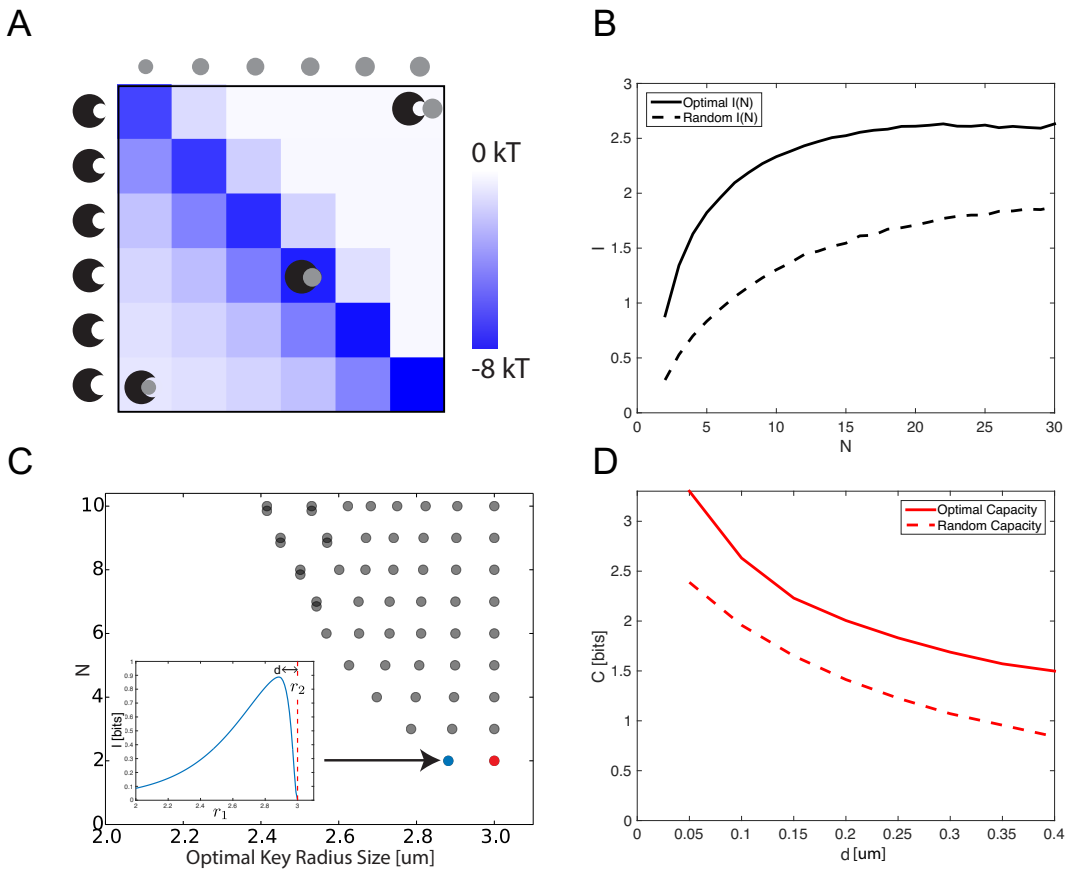
Thus far we have focused on locks and keys interacting exclusively through a single kind of physical interaction. Using our quantitative framework, we may ask how capacity increases when multiple sources of specificity, such as shape and color, are combined in a single set of locks and keys. As is known in information theory[5,32], the combined capacity of two interacting channels can be significantly higher than sum of the individual capacities.

6.5.1 Linking Two Systems

The simplest model for combining two channels is to physically link a lock of system 1 to a lock of system 2. We assume that there is no interaction between the two parts of the lock, or between the key from one system with the lock of the other system. (We do not take into account entropic effects due to avidity.) Thus for a linked system (which we write as $\text{System}_1 \oplus \text{System}_2$), the two independent systems with gaps of Δ_1 and Δ_2 are combined such that $\Delta_{Tot} = \Delta_1 + \Delta_2$. Hence the gap distribution of the linked system is the convolution of the independent systems: $\rho_{Tot}(\Delta) = \rho_1(\Delta) * \rho_2(\Delta)$, and the capacity can be computed using Eqn. 6.3 in terms of the gap distributions of the individual systems.

When two channels are linked in this form without any interaction, we expect the total capacity of the system to be $C_{Tot} = C_1 + C_2$[32]. We explicitly compute this linked capacity for the physical system shown in Fig. 6.5A (left), in which a color system of length L is linked to a shape system of length L (Shape $\oplus$ Color). The distribution $\rho_{Tot}(\Delta)$, obtained by convolving $\rho_{color}(\Delta)$ and $\rho_{shape}(\Delta)$ is shown in Fig. 6.5B. The

Figure 6.4: Pacman lock/key pairs demonstrate the capacity of shape space. **A)** Interaction energies of 6 optimally selected lock-key pairs. Cognate locks and keys fit snugly, while crosstalk is most severe between small keys and larger locks. **B)** Mutual information plotted as a function of the number of pairs N shows that both optimal (solid) and random (dashed) sets of $\{r_i\}_N$ display a maximum in mutual information. (Random pairs are drawn uniformly from [1,3] μm.) **C)** Each row, plotted at $y = N$, shows the optimal $\{r_i\}_N$ for N lock-key pairs. After saturation ($N > 6$), particles are duplicated (overlapping circles). Inset: Mutual information with r_2 held fixed at 3 μm shows how mutual information varies as r_1 changes. ($d = 100$ nm, $s_{\text{max}} = -8\ k_BT$). **D)** The capacity increases with smaller depletant particle size. On-target binding at $r = R_{max}$ is fixed to $-8\ k_BT$.



resulting capacity $C_{Tot} \approx C_1 + C_2$ is additive up to $\log L$ corrections that are small when L is large (see Appendix A).

6.5.2 Mixing Two Systems

In a mixed system, the physics of the individual systems are combined, and there is no general formula for the resulting gap distribution since $\Delta_{Tot} \neq \Delta_1 + \Delta_2$. We study a model in which shapes are coated with chemical colors, and we denote mixed systems by $\text{System}_1 \otimes \text{System}_2$ (Fig. 6.5A, right). The energy is the sum of the shape and color interactions, but the color interaction energy implicitly depends on the shape; only when the surfaces are near each other can the color-dependent interaction matter. We assume a distance dependence of the color interaction, with length scale λ_{color}, such that the energy of interaction decays as $e^{-h/\lambda_{\text{color}}}$ for two surfaces separated at a distance h.

We can intuitively understand how the mixed model differs from the linked model by examining random off-target pairs, as shown in Fig. 6.5A. In the $\oplus$ model, crosstalk arises from accidental matches in *either* independent channel; hence the crosstalk is simply the sum of the number of matching sites in the two channels. However, in the $\otimes$ model, crosstalk in the color channel can arise at a site only when there is an accidental match in *both* color and shape channels at that site. For example, in Fig. 6.5A, all three matching color sites contribute to crosstalk in the $\oplus$ model. However, in the $\otimes$ model, these three sites are not accidentally matched in the shape channel; since the three color sites are not in contact, they do not contribute to crosstalk. As a result, off-target binding is generally weaker and the typical gap Δ higher in the $\otimes$ model, as we find in Fig. 6.5B. Thus the mixing of shape and color in this interactive manner increases the capacity.

We may further examine how the capacity changes as a function of λ_{color}, the inter-

action range of the color system. When λ_{color} is small compared to δ, the maximum height of local shape features, shape features can be easily distinguished by the color force and so the color and shape work in concert to increase capacity. Increasing λ_{color} blurs the shape contours and the color interactions no longer distinguish shapes, thereby becoming less specific. Indeed, Fig 6.5C shows that when $\lambda_{\text{color}}/\delta$ becomes large, the color system and the shape system act independently, and the capacity relaxes to the capacity of the linked system Shape $\oplus$ Color.

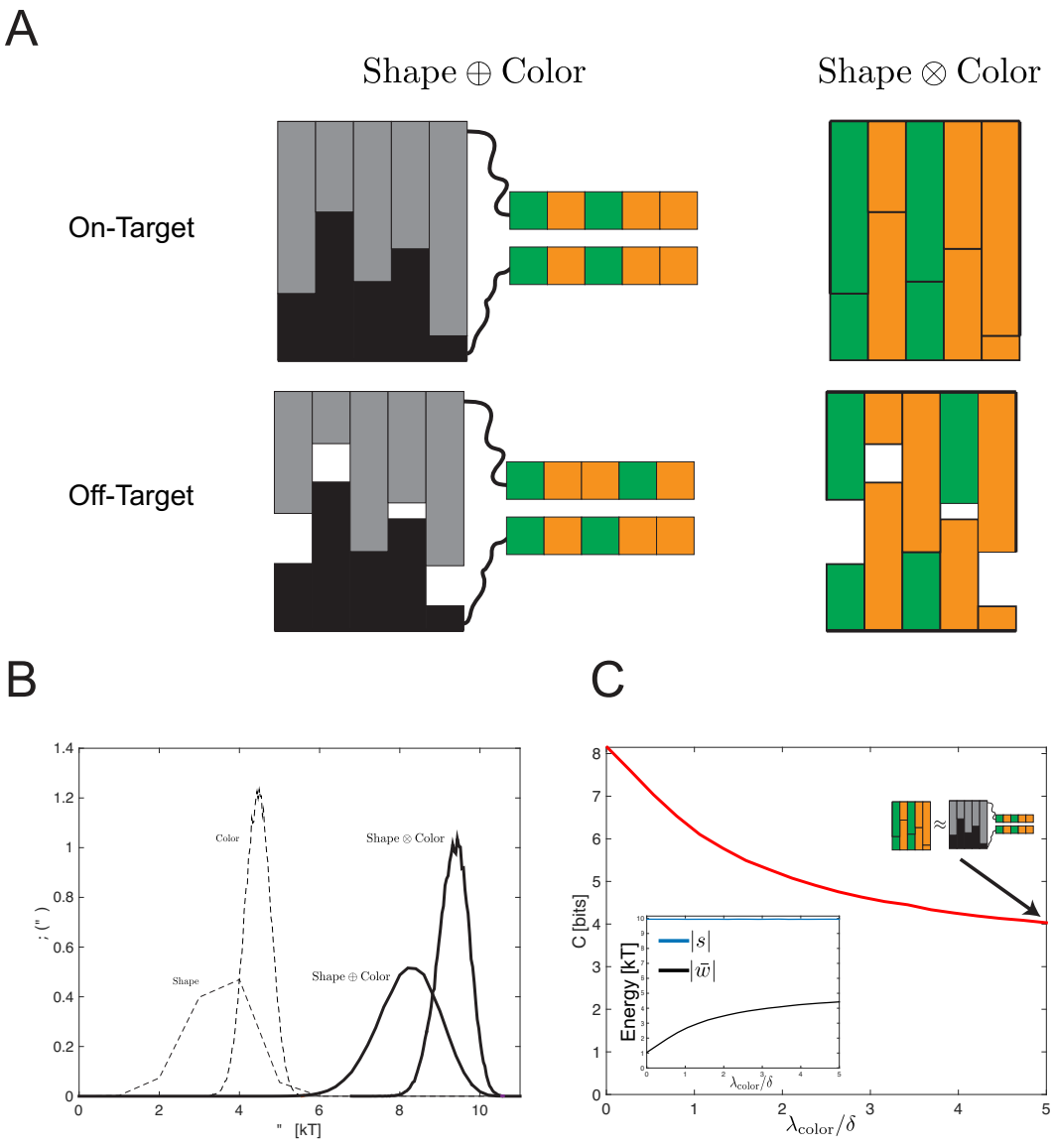
In summary, laying out color-based codes on undulating surfaces significantly reduces the total crosstalk since color-matched sites must also be matched in shape to contribute to crosstalk. Such color-shape synergy persists so long as the spatial range of color interactions is shorter than the length scale of shape variation.

6.6 Discussion

Here we have shown that mutual information provides a general metric for specificity, bounding the number of distinct lock-key pairs that can be supported by systems of programmable specific affinities. Mutual information is well suited as a measure of specificity for many reasons. First, mutual information is a global measure of specificity, accounting for all possible interactions between N species of locks and keys. Second, as a result, it provides a precise answer as to how many particle pairs can be productively used in a given system. As N is increased, crosstalk necessarily increases as we crowd the space of possible components (Fig. 6.1B,[113]) with more and more lock-key pairs. Capacity is determined by the point $N = N_C$ at which the information gain due to larger N is negated by the increase in crosstalk.

Third, we can use mutual information to quantitatively compare disparate types of programmable interactions, from DNA hybridization to depletion driven interactions.

Figure 6.5: Combining Color and Shape. **A)** Shape and color can be 'linked' in an independent manner (left, $\oplus$), or 'mixed' in a dependent manner (right, $\otimes$), where the shapes are coated with the chemical binding agent. Crosstalk in $\oplus$ results from accidental matches in *either* channel, while in the $\otimes$ model, accidental matches in color contribute to crosstalk only if shapes are also matched at the same sites. **B)** As a result, the gap energy for Shape $\otimes$ Color is higher than for $\oplus$. Here $L = 10$ for both color and shape, $d = 0.05$ μm, $\delta = 1$ μm, $A = 4$, and $\lambda_{\text{color}} = 0.01$ μm. **C)** The capacity as a function of the spatial range λ_{color} of color-based interactions. When λ_{color} is smaller than δ, the typical size of shape-based features, shape helps reduce crosstalk in color. This synergy is lost when $\lambda_{\text{color}} \sim \delta$, and color and shape act independently. ($\epsilon_{x^l=y^l} = 0.5\ k_BT, \epsilon_{x^l\neq y^l} = 0.25\ k_BT$, see Appendix A).



Our framework can also quantitatively predict how varying physical parameters (e.g., depletion particle size, range of interactions, elastic modulus of shapes) raises or lowers specificity. The models we discuss can be further refined in various ways, for example by allowing DNA strands to fold, examining shapes in three dimensions, or taking into account the entropic effects of multivalency and avidity[73].

Using such an approach, we found that (1) shape complementarity intrinsically suffers less crosstalk than 'color' (i.e., chemical specificity)-based interactions and (2) multiple physical interactions, such as color-based and shape-based interactions, can be combined in a synergistic manner, giving a capacity that is greater than the sum of the parts. Such predictions are especially valuable, given the proliferation of different mechanisms for creating and combining distinct mechanisms of specificity: mutual information provides an unbiased way of comparing their efficacy to each other. As programmable specificity continues to drive technological developments in self-assembly[154], understanding how the mutual information of paired components can be built up towards creating larger, multi-component objects is a critical future direction of this work.

While we focus on applications to colloidal systems, we note that the framework developed here can be employed to study biological systems as well. In 1890, Emil Fischer proposed the 'lock and key' model as an analogy for understanding enzyme-substrate specificity[45], focusing on the physical shapes of paired interacting components; mutual information encompasses this idea and can be applicable to a large number of biological systems. In particular, our model is useful for predicting the differences between interacting proteins that use shape complementarity alone and those that combine both shape and electrostatic complementarity (e.g. Dpr-DIP vs Dscam proteins[23]), and may also be applied to a host of other biological interaction networks[70] where information transmission and pair specificity play critical roles in biological function (e.g. HKRR

proteins[117] and the immune system[113]). Crucially, the mutual information model provided above is flexible enough to be extended to some of the challenging physics encountered in biology. Nonequilibrium systems can be accounted for by computing time dependent probabilities of interactions instead of the equilibrium probabilities, while hypotheses for increased specificity like 'induced fit' and recent variants[125] can be tested directly for their impact on capacity.

In this work, we have shown that mutual information is a powerful tool to describe diverse specificity models. The strength of our framework is that it is broadly applicable - it may immediately be applied to any system for which the pairwise energies of interactions are known, in both biology and in synthetic experiments. We believe that using the capacity as a measure of system specificity will provide a simple metric for analyzing, comparing, and optimizing systems of programmable interactions.

7

A Random Matrix Theory Perspective of Pragmatic Principal Component Analysis

7.1 Introduction

Principal Component Analysis (PCA) is widely used as an unsupervised dimensionality reduction tool. The goal is to discover the relationship between variables by measuring correlations across many samples. The theoretical underpinning of PCA is provided by random matrix theory, which predicts how the eigenvectors of the sample covariance matrix (principal components) are affected by signal. When there are many more samples than variables, the eigenvectors of the sample covariance matrix with the largest eigen-

values accurately reveal dependencies between the variables. However, if the ratio of the number of variables to samples is too large, even the top principal component retains no information about the true signal[8,13]. This increasingly presents an issue as large numbers of variables are now routinely measured simultaneously in experiments[114,29].

To circumvent these issues, researchers have introduced 'pragmatic' modifications to PCA, wherein the data matrix is transformed in an application specific manner before PCA is used. We discuss three primary categories of prior knowledge. 1) Knowledge of imbalanced variables, for which there is prior information that some variables are more statistically important than others. A prominent recent example is Statistical Coupling Analysis (SCA), which performs PCA after first reweighting the variables in a protein sequence alignment to highlight conserved residues[57]. 2) Knowledge of imbalanced samples, for which there is prior knowledge that the samples are not identically or independently distributed. This often arises when samples are processed in batches, leading to within-batch correlations. These correlations can have nontrivial effects on the analysis, as acutely illustrated in a recent controversy[146] over analysis of gene expression data in an ENCODE study[84]. In this case, the pragmatic PCA approach, which corrected for batch effects, led to the opposite conclusions compared to naive PCA[53]. 3) Knowledge of signal direction, for which there is information regarding nontrivial correlations between variables. For example, signal recovery in stock data is often enhanced by averaging stocks in the same sector (e.g. technology, transportation, etc.) as they are known to be highly correlated. Similarly a priori knowledge of the correlations between genes can enhance signal recovery in transcriptional networks.

Up until now, the validity of these 'pragmatic' modifications has been determined *ex post facto*; methods are argued to work because they give more interpretable results. The goal of this work is to provide a mathematical framework that elucidates

when pragmatic PCA modifications are appropriate. We focus on the fundamental detectability threshold of random matrix theory, which delineates when a signal contained in n independent observations of p variables can be distinguished from noise[8]. The largest principal components will only have have significant projection onto the subspace spanned by the signal variables if the signal strength exceeds $\sigma\sqrt{p/n}$, where σ is the standard deviation of the noise[87,71,43,7,8,111]. Otherwise, there will be no signal present in the eigenvectors.

Here, we calculate the conditions under which linear transformations of the data that incorporate prior information improve this signal detection threshold. With increasing accuracy of the prior, we show that both the signal detection threshold and the alignment of the top principal component with the true signal improve. On the other hand, caution must be exercised, in that there are parameter regimes where the detected signal results solely from the imposed linear transformation of the data, i.e. the prior information. This chapter is organized as follows. We first review the major results of random matrix theory relating to PCA. Then, we analyze models of pragmatic PCA for the three categories, and illustrate through analysis and simulations how they affect the signal detection threshold. We then apply our framework to a case study of pragmatic PCA, Statistical Coupling Analysis, and demonstrate how this procedure can be understood using random matrix theory.

7.2 Random Matrix Theory and PCA

Given a mean-centered p by n data matrix X, containing n independent samples of p variables, PCA computes the eigenvalues and eigenvectors of the sample covariance matrix $E = n^{-1}XX^T$. The basic hope of PCA is that a small number of the eigenvectors of E, which correspond to the largest eigenvalues of E, will accurately describe the signal

in X. Here we review how random matrix theory rigorously defines when this procedure will succeed. To see this, we will analyze the secular equation of the simplest model, the rank-one 'spiked' covariance model[71]. Let the entries of X be drawn from a mean-zero Gaussian distribution with a p-by-p covariance matrix $\Sigma = \mathbb{I}_p + s\mathbf{e_s}\mathbf{e_s}^T$: the signal is a rank-one perturbation in the direction of $\mathbf{e_s}$ with signal strength s where $s \geq 0$.

If $s = 0$, or equivalently the variables in X are completely uncorrelated, the eigenvalue spectrum of E is known to follow the well studied Marčenko-Pastur (MP) distribution[87], compactly supported between $\mu_\mp = (1 \mp \sqrt{p/n})^2$, while the eigenvectors of E will be random and normally distributed on the p-dimensional sphere. When $s \neq 0$, we may calculate the spectrum by noting that $E = n^{-1}XX^T = n^{-1}\Sigma^{1/2}GG^T\Sigma^{1/2}$ is similar to, and therefore has the same eigenvalues as, $M = n^{-1}GG^T\Sigma = n^{-1}GG^T(\mathbb{I} + s\mathbf{e_s}\mathbf{e_s}^T)$. The eigenvalues of M obey the equation $|\lambda\mathbb{I} - M| = 0$, which can be rewritten[13] as

$$|\lambda\mathbb{I} - n^{-1}G^TG||\mathbb{I} - (\lambda\mathbb{I} - n^{-1}G^TG)^{-1}n^{-1}G^TGs\mathbf{e_s}\mathbf{e_s}^T| = 0 \tag{7.1}$$

Thus, either λ is an eigenvalue of $n^{-1}G^TG$, and so contained within the MP distribution bulk, or 1 is an eigenvalue of the rank one matrix $\Psi \coloneqq (\lambda\mathbb{I} - n^{-1}G^TG)^{-1}n^{-1}G^TGs\mathbf{e_s}\mathbf{e_s}^T$. Diagonalizing $n^{-1}G^TG = \sum_i \mu_i\mathbf{v_i}\mathbf{v_i}^T$ we can rewrite Ψ as

$$\Psi = s\sum_{i=1}^{p}\frac{\mu_i}{\lambda - \mu_i}(\mathbf{v_i}^T\mathbf{e_s})^2\mathbf{v_i}\mathbf{v_i}^T \approx \frac{s}{p}\sum_{i=1}^{p}\frac{\mu_i}{\lambda - \mu_i}\mathbf{v_i}\mathbf{v_i}^T \tag{7.2}$$

since $\mathbf{e_s}$ is approximately random with respect to the basis of random eigenvectors $\mathbf{v_i}$. Because Ψ is rank-one, the condition for λ to be larger than the upper limit of the MP

bulk (denoted μ_+) is $Tr(\Psi) = 1$, or

$$1/s = \frac{1}{p}\sum_{i=1}^{p}\frac{\mu_i}{\lambda - \mu_i} \xrightarrow[p\to\infty]{} \int_{\mu_-}^{\mu_+}\frac{\mu}{\lambda-\mu}dF(\mu). \tag{7.3}$$

where $F(\mu)$ is the MP distribution function. The final expression defines the T-transform of $n^{-1}G^TG$, a monotonically decreasing function of $\lambda > \mu_+$ which attains its largest value $T_{max} = \sqrt{n/p}$ in the limit of $\lambda \downarrow \mu_+$ [100,8,19]. If $1/s > T_{max}$, there is no λ which can satisfy Eqn. Eqn. (7.3), providing a sharp detectability threshold for the signal strength s, known as the BBP phase transition: if $s < \sqrt{p/n}$, the largest eigenvalue of E, denoted λ_1, will remain in the MP bulk. The value of λ_1 when s is above the BBP threshold can be analytically computed from the T-transform of $n^{-1}G^TG$, which gives

$$\lambda_1 \to \begin{cases} \left(1+\sqrt{\frac{p}{n}}\right)^2 & \text{if } s < \sqrt{p/n} \\ (s+1)(1+\frac{p/n}{s}) & \text{if } s > \sqrt{p/n} \end{cases} \tag{7.4}$$

A similar argument shows that the first principal component ($\mathbf{v_1}$) has non-trivial alignment with $\mathbf{e_s}$ only when s is above the detectability threshold:

$$f^2 \coloneqq |\mathbf{v_1}^T\mathbf{e_s}|^2 \to \begin{cases} 0 & \text{if } s < \sqrt{p/n} \\ \left(1-\frac{p/n}{s^2}\right)\left(1+\frac{p/n}{s}\right)^{-1} & \text{if } s > \sqrt{p/n} \end{cases} \tag{7.5}$$

These expressions quantify how well PCA performs at signal identification.

7.3 Pragmatic PCA and the Signal Detection Threshold

We now examine how the signal detection threshold and signal alignment (Eqns. (7.4) & (7.5)) are altered by pragmatic PCA. By pragmatic PCA, we refer to application-specific modifications of X that incorporate prior information. (We do not consider

situations where data processing does not explicitly involve prior information, such as transformations to give the data zero mean and unit variance, or logarithmically transforming the data.) In what follows, we analyze mathematical models for three basic classes of pragmatic PCA, and determine how signal detection is changed, depending on the quality of the prior information.

Category 1: Imbalanced Variables

Due to experimental limitations, it is not possible to measure every possible variable. Moreover, statistical power is sacrificed as the number of variables p is increased, as they are measured by a limited number of n samples. Thus in practice, experiments are designed such that only variables that are thought to be relevant are measured, with the aim of avoiding nuisance parameters. Even after such choices are made, some pragmatic PCA techniques, such as SCA, reweight variables according to prior information about their significance, skewing which variables are detected in signal identification.

This informal process can be formalized with an intuitive linear transformation, where prior information about variable importance is used to delete variables from a data matrix measured on a larger set of variables. Since the signal strength must be larger than the BBP threshold $\sqrt{p/n}$ to be detected, it is temping to randomly delete variables in order to decrease p. However, this process actually hurts detectability[18], since it also decreases the signal strength. In contrast, we show that detectability can improve if one deletes variables using prior information about which variables carry the signal.

Consider the data matrix X described above, and assume that $\mathbf{e_s}$, the signal vector, has B non-zero entries grouped as a block. To delete variables, define $\tilde{X} = JX$, where J is a diagonal matrix with αp ones at the rows we wish to keep and $(1-\alpha)p$ zeros at the rows we wish to delete ($0 \leq \alpha \leq 1$). We examine how the spectrum and eigenvectors

of $\tilde{E} = \tilde{X}\tilde{X}^T/n$ differ from that of $E = XX^T/n$: does this pragmatic PCA technique improve signal detection?

After deletion, $p \to \tilde{p} = \alpha p$. We characterize the quality of the prior by b ($0 \leq b \leq 1$), where bB is the number of rows from the signal block that remain after the deletion procedure, so $s \to \tilde{s} = bs$. Thus the detectability requirement becomes

$$\tilde{s} > \sqrt{\tilde{p}/n} \implies bs > \sqrt{\alpha p/n}. \tag{7.6}$$

Eqn. (7.6) shows the prior acts on detectability in two different ways. On the one hand, it degrades the signal $s \to bs$, since some of the row deletions (incorrectly) remove signal. On the other hand, it lowers the noise by reducing the number of variables $p \to \alpha p$. If the entries of J are chosen randomly ($b \approx \alpha$), the deletion procedure makes the detectability threshold more stringent[18], demanding $\sqrt{\alpha}s > \sqrt{p/n}$. But if the prior is accurate, with $b > \sqrt{\alpha}$, then deletion increases the proportion of rows in the signal block, relaxing the detectability requirement. Thus for large b a signal may go from being undetectable in X to detectable in $\tilde{X}$, showing how prior information can greatly aid in signal recovery.

When the signal is above the detectability threshold, the overlap of $\tilde{\mathbf{v}}_\mathbf{1}$ (the largest principal component of $\tilde{X}$) with the signal $\mathbf{e_s}$ can be obtained from Eqn. (7.5) using $\tilde{s}$ and $\tilde{p}$:

$$\tilde{f}^2 = |\tilde{\mathbf{v}}_\mathbf{1}^T \mathbf{e_s}|^2 = \left(1 - \frac{\alpha p/n}{b^2 s^2}\right)\left(1 + \frac{\alpha p/n}{bs}\right)^{-1} \tag{7.7}$$

Thus the signal recovery depends crucially on b, the quality of the prior information. This agrees with simulations (Fig. 7.1A), which examine the case of a signal which is below the BBP threshold and therefore undetectable (in the infinite n limit) in X while detectable in $\tilde{X}$ for large b.

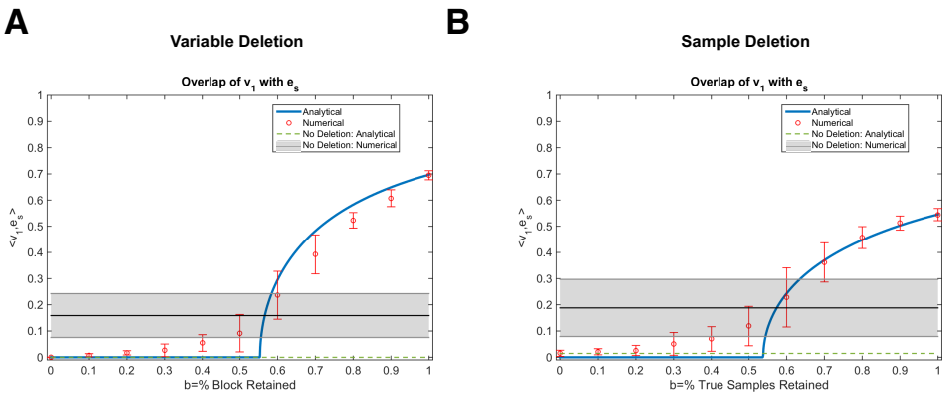


Figure 7.1: Deletion of variables or samples recapitulates theory. **A** Deletion of variables is simulated (red) and matches the analytical prediction (blue) for the overlap of the top eigenvector with the signal $\mathbf{e_s}$, as a function of b, the number of variables that are correctly retained from the block. Here, signal recovery is very limited without prior information because the data is below the BBP threshold, and so the expected overlap between the top eigenvector and the signal, when no deletion is performed, is zero. Due to finite p, n, however, the recovery of the signal without deletion is finite (black: mean, gray: standard deviation). Here $p = 6000$, $n = 3000$, $\alpha = .3$, $B = 1800$; for each point, the mean and standard deviation over 20 trials is shown. **B** Deletion of samples (in this case, posited outliers), and the predicted overlap between the top eigenvector and the signal vector after samples are deleted is shown in blue. Overlap is plotted as a function of b, the percent of non-outlier samples that are retained after deletion. Red: Mean and standard deviation of simulations, averaged over 20 trials. Without deletion, the expected overlap is a small but finite number, as the system is above the BBP threshold. Black/Gray: Average/Standard Deviation of the overlap without deletion is above the expected value due to finite p effects. Here $p = 4000$, $n = 2000$, $n_1 = 600$, $\alpha = .3$, and $s = 4.8$.

Category 2: Imbalanced Samples

A second category of prior information relates to knowledge of non-independence or non-identicality of samples. For example, in genomic analyses, the genetic sequence at many sites (variables) is compared across many individuals (samples), yet since individuals are related by descent, the samples are not independent. Another common example is batch effects: in experiments samples may be process in batches, so that samples processed in the same batch are more correlated. If these issues are not controlled for, PCA can give very misleading results. Hence different forms of pragmatic PCA have been developed that employ prior information to correct the non-i.i.d. nature of a dataset.

Here we discuss one common correction technique, the deletion of outlier samples – samples that are drawn from a different distribution than the majority[27]. Consider a Gaussian mixture model where samples are drawn from two distributions. Suppose there are n_1 samples drawn from the 'correct' distribution $N(0, \Sigma)$, arranged in a p by n_1 matrix X_1, and $n_2 = n - n_1$ samples drawn from a 'corrupted' distribution $N(0, \Sigma_c)$, organized as a p by n_2 matrix X_2. The observed dataset is a column permutation of the concatenation $[X_1, X_2]$, with the identity of each sample unknown. The addition of the outlier points X_2 to the dataset hampers the ability to recover Σ. For example, if $\Sigma = \mathbb{I} + s\mathbf{e_s}\mathbf{e_s}^T$ and $\Sigma_c = \mathbb{I}$, then the sample covariance matrix of the mixture will indistinguishable from the sample covariance matrix computed from samples drawn from a distribution with covariance matrix $\mathbb{I} + s(n_1/n)\mathbf{e_s}\mathbf{e_s}^T$, and total signal strength $s(n_1/n)$. Thus, when outliers are mixed in, the threshold for detectability is $s(n_1/n) > \sqrt{p/n}$.

Outlier deletion can be formulated as performing PCA on the altered matrix $\tilde{X} = XJ$, where J is an n by n diagonal matrix with ones at locations where samples should be

kept, and zeros at the locations of samples which should be removed. The performance of such a procedure depends on the quality of prior knowledge about the sample identities. If the deletion procedure leaves αn samples, where $0 \leq \alpha \leq 1$, and bn_1 of the uncorrupted samples remain after deletion ($0 \leq b \leq 1$), then $n_1 \to \tilde{n_1} = bn_1$ and $n \to \tilde{n} = \alpha n$. The largest eigenvalue of $\frac{1}{\alpha n}\tilde{X}\tilde{X}^T$ will exit the bulk only if $s\frac{bn_1}{\alpha n} > \sqrt{p/(\alpha n)}$. This is the analogue of Eqn. (7.6) for row deletion, and shows that outlier deletion can increase or decrease the detectability threshold, depending on b and α. Furthermore, as before, we can calculate $\tilde{f}$ to quantify how well the principal eigenvector $\tilde{\mathbf{v}}_\mathbf{1}$ reconstructs $\mathbf{e_s}$ using Eqn. (7.5), replacing s with $\tilde{s} = s\frac{bn_1}{\alpha n}$ and n with $\tilde{n} = \alpha n$. These results agree with simulations (Fig. 7.1 B) and are intuitive in nature: deletion of corrupted samples cam improve signal reconstruction, but only when there is high quality prior information regarding which samples are corrupted. Incorrect prior information will decrease signal detectability.

Category 3: Signal Direction

Frequently in experiments there is preexisting knowledge, perhaps from orthogonal experiments, regarding correlations between variables. For example, in gene expression datasets known genetic pathways give prior information about co-expressed genes, while in financial data, industry sectors gives prior information about the companies with correlated stocks.

To understand the conditions where such knowledge can improve signal detectability, we again consider the model where the data X has a 'spiked' covariance matrix $\Sigma = \mathbb{I}_p + s\mathbf{e_s}\mathbf{e_s}^T$. Suppose we have a guess for the signal direction, $\mathbf{e_g}$, and we modulate our confidence in that guess with a parameter $s_g > 0$. Let $b \coloneqq |\mathbf{e_g}^T\mathbf{e_s}|$ measure the alignment between the guess vector and the signal vector. How does signal detectability

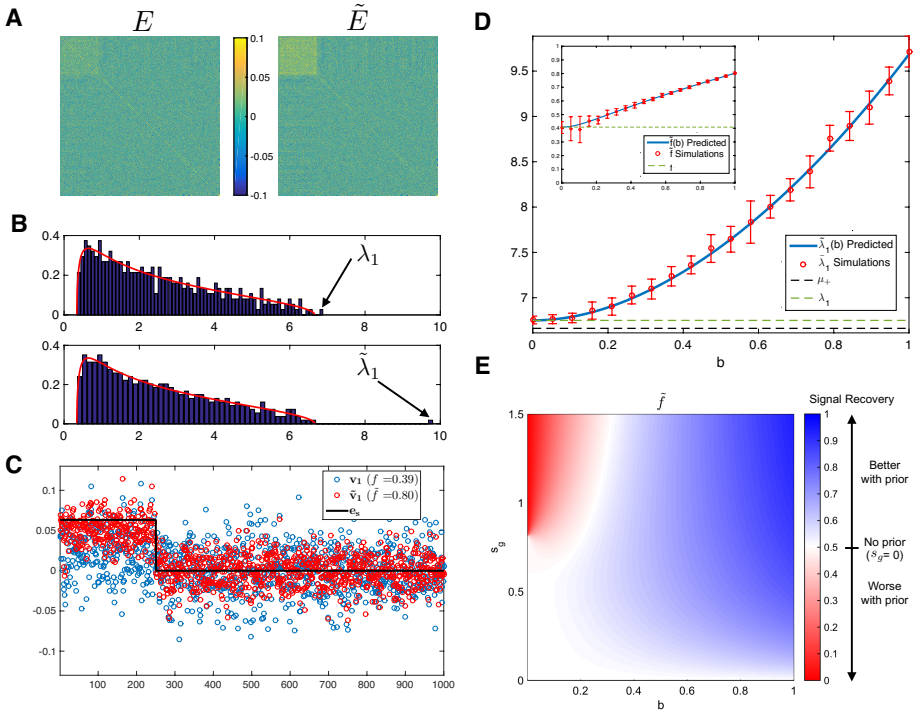


Figure 7.2: Signal Recovery Above BBP. **A-C** An illustrative example demonstrates how prior information improves signal detection ($p = 1000, n = 400, s = 2.2, s_g = 0.5, b = 0.95$). **A** The signal block in $\tilde{E}$ is enhanced relative to E. **B** Top: The bulk of the eigenvalues of E conform to the MP distribution (red), and the largest eigenvalue λ_1 is outside the bulk. Bottom: Using prior information, $\tilde{\lambda}_1$ is pushed further out of the bulk. **C** Alignment of $\mathbf{v_1}$ (blue) with $\mathbf{e_s}$ (black) is weaker than alignment of $\tilde{\mathbf{v}}_\mathbf{1}$ (red). **D** Predictions (blue) and simulations (red) for $\tilde{\lambda}_1$ show it increases relative to λ_1 (green) and the MP edge (black) as b increases. Error bars show standard deviation over 20 runs. **D, Inset** Predictions (blue) and simulations (red) for the alignment $\tilde{f} = \tilde{\mathbf{v}}_\mathbf{1}^T \mathbf{e_s}$, compared with the original alignment f; alignment with the signal improves with increasing b. (In D, $p = 5000, n = 2000, s = 2, s_g = 0.5$, and $\mathbf{e_s}$ was chosen randomly for each run). **E** Signal alignment $\tilde{f}$ for different parameters of b (prior quality) and s_g (prior strength). Without using prior information, $f = 0.5$. $\tilde{f}$ can be larger than this if b is sufficiently large and s_g is given sufficient weight. However, if b is small, and s_g is large, the signal alignment will be worse. ($p = 1000$, $n = 400$, $s = 2.2894$)

changes as function of the quality b of the guess, the strength of the original signal (s), and the strength of our guess (s_g)?

Define the guess covariance matrix $J = \mathbb{I}_p + s_g \mathbf{e_g} \mathbf{e_g}^T$, and consider $\tilde{X} = JX$, such that

$$\tilde{X} = (\mathbb{I} + s_g \mathbf{e_g} \mathbf{e_g}^T)(\mathbb{I}_p + s \mathbf{e_s} \mathbf{e_s}^T)^{1/2} G \tag{7.8}$$

where we have used $X = \Sigma^{1/2} G$. To evaluate eigenvalues and eigenvectors of the data matrix $\tilde{E} = \tilde{X}\tilde{X}^T/n$, we may think of $\tilde{X}$ as n draws from a p-dimensional Gaussian with correlation matrix $\tilde{\Sigma}$:

$$\tilde{\Sigma} = \mathbb{I} + \tilde{s} \mathbf{e_s} \mathbf{e_s}^T + s_\perp \mathbf{e}_\perp \mathbf{e}_\perp^T + s_{\text{cross}} (\mathbf{e}_\perp \mathbf{e_s}^T + \mathbf{e_s} \mathbf{e}_\perp^T)$$

Here $\mathbf{e}_\perp$ is the unit vector orthogonal to the signal vector defined by the guess vector, $\mathbf{e}_\perp = (\mathbf{e_g} - b\mathbf{e_s})(1-b^2)^{-1/2}$, and $\tilde{s}$, $s_\perp$ and s_{mix} are all scalars that modulate the weight in the two directions and in their crossterms. In general the first eigenvector of $\tilde{\Sigma}$ is a combination of $\mathbf{e_s}$ and $\mathbf{e}_\perp$, but we may examine two simple limits: (1) When $b = 1$ (completely accurate guess), then $s_\perp = s_{\text{mix}} = 0$ and $\tilde{s} > s$, so the procedure amplifies the signal vector. (2) When $b = 0$ (completely inaccurate guess), $\tilde{\Sigma}$ retains signal s in the direction of $\mathbf{e_s}$ and gains signal strength $s_\perp > 0$ in the direction orthogonal to the signal (and $s_{\text{mix}} = 0$). If s_g is large enough, $s_\perp > \tilde{s}$, and the first eigenvector of $\tilde{\Sigma}$ has zero overlap with $\mathbf{e_s}$.

To gain intuition for how this procedure improves signal reconstruction, Fig. 7.2A shows a sample covariance matrix $E = XX^T/n$ of data whose signal vector $\mathbf{e_s}$ has ones in the first 250 variables and zeros elsewhere; note that one can barely identify the block in the upper left corner. The top eigenvalue λ_1 is separated from the bulk close to its expected value of $(s+1)(1+\frac{p/n}{s}) = 6.84$, and $\mathbf{v_1}$, while noisy, has some overlap with $\mathbf{e_s}$

(f=0.39). We choose a near perfect guess vector $\mathbf{e_g}$ with $|\mathbf{e_g}^T\mathbf{e_s}| = 0.95$, and construct $\tilde{X}$ as outlined above. Fig. 7.2A shows that $\tilde{E}$ now has a highly visible signal block. Its top eigenvalue $\tilde{\lambda}_1$ is much more separated from the bulk (Fig. 7.2B), and $\tilde{\mathbf{v}}_\mathbf{1}$ has twice as much overlap with $\mathbf{e_s}$, with $\tilde{f} = 0.80$ (Fig. 7.2C).

More generally, the exact expressions for $\tilde{X}$, given by the analogs of Eqn. (7.4) and Eqn. (7.5), show that a high quality prior information improves performance. Fig. 7.2D shows the analytical and numerical results when the original signal strength is above the detectability threshold, i.e. $s > \sqrt{p/n}$ such that λ_1 is separated from the bulk: $\tilde{\lambda}_1$ increases above λ_1 as b increases, and likewise the alignment of the top eigenvector with the signal $\tilde{f}$ increases above f with increasing b. However, signal detectability can degrade with the prior if b is small and s_g is large. Fig. 7.2E plots $\tilde{f}$ as a function of (b, s_g). For small b and large s_g, $\tilde{f} < f$. This is intuitive, as an incorrect prior that is given a lot of weight should give poor results. We find similar results when the untransformed signal is below the BBP threshold ($s < \sqrt{p/n}$). Fig. 7.3A shows the spectrum from the same model as Fig. 7.2B, except with s below the phase transition. PCA on X recovers no information: λ_1 is near its expected value of μ_+ at the edge of the bulk, and the signal alignment is correspondingly very small ($f = 0.03$). With an accurate guess ($b = 0.95$), PCA shows that $\tilde{\lambda}_1$ is pushed outside the bulk and correspondingly the signal alignment is significantly increased ($\tilde{f} = 0.70$). This is a dramatic demonstration of the effect of prior information.

However, the improvement on $\tilde{f}$ may not always be as rewarding. Fig. 7.3B shows $\tilde{f}$ as a function of b and s_g: with all other parameters held fixed, b must exceed a critical threshold b_c before the eigenvector has nonzero overlap with the signal. For $b < b_c$, $\tilde{f} = 0$ (in the infinite limit). Note that when s is small such that E is below the phase transition, $|\mathbf{v_1}^T\mathbf{e_s}| = 0$, so using prior information never results in a $\tilde{\mathbf{v}}_\mathbf{1}$ which has

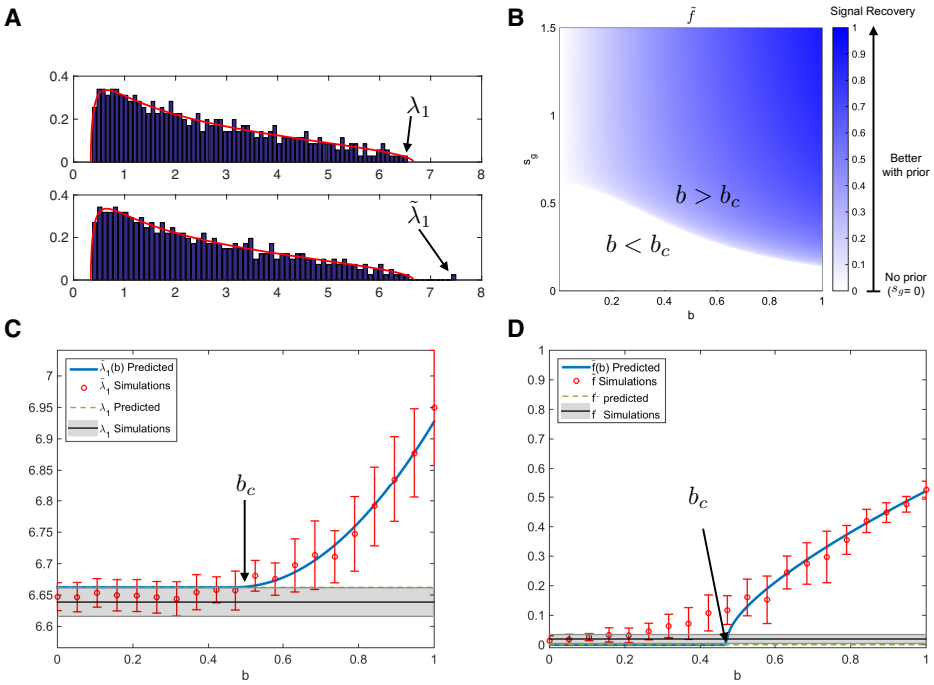


Figure 7.3: Signal recovery below BBP. **A** An example dataset. Top: When $s < \sqrt{p/n}$, PCA performed on X results in negligible alignment of the largest eigenvector with the signal ($f = 0.03$), and the largest eigenvalue λ_1 is located at the edge of the MP spectrum ($\lambda_1 \approx \mu_+$). Bottom: When PCA is performed on $\tilde{X}$ with high quality prior information, the largest eigenvalue gets pushed out of the bulk and the signal alignment is dramatically improved ($\tilde{f} = 0.70$). ($p = 1000, n = 400, s = 1, s_g = 0.5, b = 0.95$.) **B** Signal alignment $\tilde{f}$ for different parameters of b (prior quality) and s_g (prior strength). Without using prior information, $f = 0$, and so $\tilde{f}$ can only improve on this. However, b must be larger than some critical value b_c in order to affect the alignment. ($p = 1000$, $n = 400$, $s = 1$). **C** Analytical (blue) and numerical (red) results for $\tilde{\lambda}_1$ show that the eigenvalue only exits the MP bulk for $b > b_c$. Since s is below the BBP threshold, the predicted λ_1 (green) is at the MP edge, but numerical simulations at finite n show a systematic downward bias (black: mean, gray: standard deviation). **C** Analytical (blue) and numerical (red) results for $\tilde{f}$ show that the overlap increases dramatically at $b > b_c$. The numerical phase transition is not as sharp as predicted; simulations show that f (black: mean, gray: standard deviation) is skewed above its expected value of zero (green). (In C and D, $p = 5000, n = 2000, s = 0.5, s_g = 0.5$, and $\mathbf{e_s}$ was chosen randomly for each run.)

worse alignment than $\mathbf{v_1}$. Analytical and numerical values for $\tilde{\lambda}_1$ and $\tilde{f}$ are shown in Fig. 7.3C,D as a function of b for a particular choice of parameters. $\tilde{\lambda}_1$ only exits the bulk and $\tilde{f}$ is only expected to be nonzero for $b > b_c$.

As an aside, we note that a high quality guess (b close to 1) is not the only way for $\tilde{\lambda}_1$ to increase relative to λ_1. If $\mathbf{e_g}$ is highly aligned with $\mathbf{v_1}$, this increases the largest eigenvalue regardless of alignment with $\mathbf{e_s}$. Alternatively, any poor guess that is given an excessive amount of weight (s_g large) will eventually dominate the spectrum. Thus, an increase of $\tilde{\lambda}_1$ relative to λ_1 should not be taken as proof for a successful use of prior information without further inspection.

7.4 Protein Sequence Analysis

Finally, we show how random matrix theory can be used to analyze a striking recent example of pragmatic PCA . Statistical Coupling Analysis (SCA) analyzes the evolutionary variation in a protein sequence alignment to understand which sequence positions control different aspects of protein function. For a protein of interest, the data matrix X contains columns representing the positions along a protein's sequence, and rows representing different versions of this protein sequence observed across evolution. X_{ij} is set to 0 or 1 depending on whether the jth sequence has the most common amino acid at position i.

SCA[57,86,28,138] performs PCA on the transformed matrix $\tilde{X} = JX$ which incorporates prior knowledge that sequence positions that are highly conserved across evolution are more likely to be important for protein function. Here J is a diagonal matrix upweights conserved positions, with $J_{ii} = \log\left[p_i^a(1-q^a)/(1-p_i^a)q^a\right]$, where p_i^a is the frequency of the most common amino acid a in row i and q^a is the background frequency of amino acid a across all proteins.

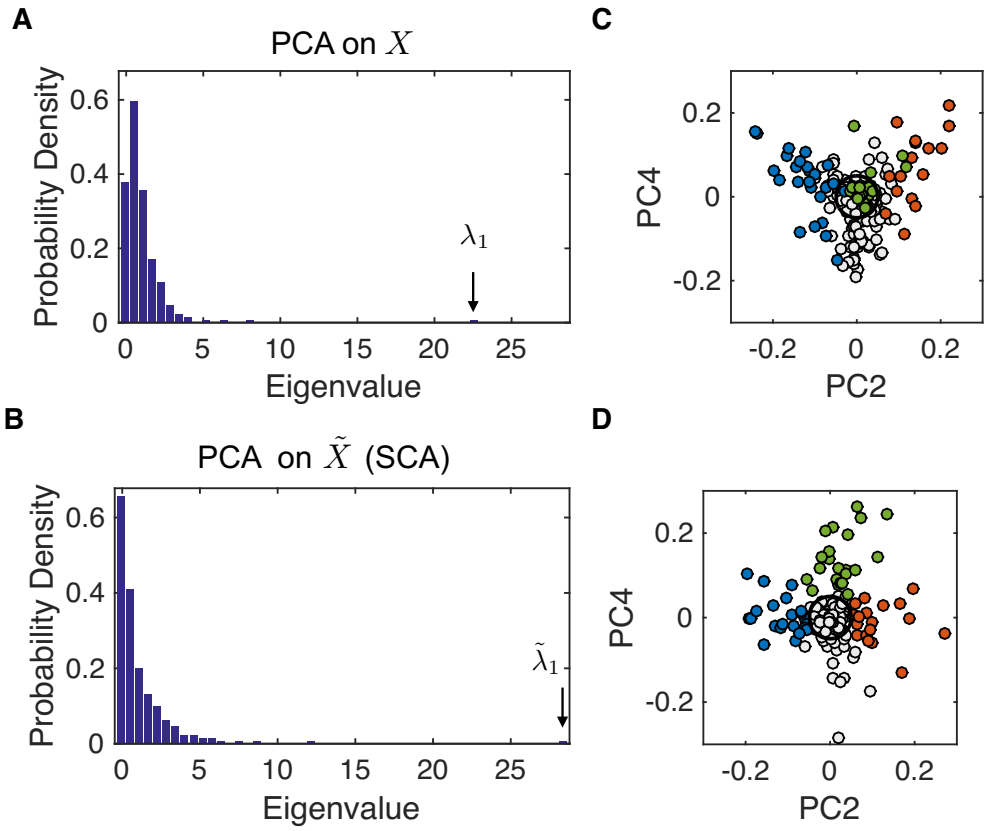


Figure 7.4: The eigenvalue spectrum of the protein sequence matrix. $\mathbf{A}$ without and $\mathbf{B}$ with prior information (arrow points to largest eigenvalue). Scatter plots of the resulting principal components 2 and 4 $\mathbf{C}$ without and $\mathbf{D}$ with prior information. The location of in D were used in[57] to identify the sectors (red, green, blue); note that in D, without prior information, the green sector is no longer detected.

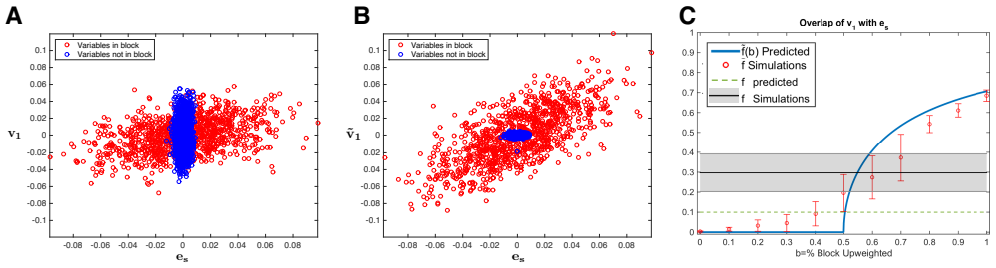


Figure 7.5: Model of SCA. The first eigenvector of **A** XX^T/n (no weighting) and of **B** $\tilde{X}\tilde{X}^T/n$ (with weighting, for $b = 1$). Both are plotted against the signal vector $\mathbf{e_s}$, which contains a block of $B = 1000$ significant variables (red) and 2000 insignificant variables (blue). The overlap of the eigenvector with $\mathbf{e_s}$ is improved and has fewer insignificant variables contributing to the eigenvector direction when weighting is performed with proper knowledge of the variable imbalance. **C** The overlap of the eigenvector with $\mathbf{e_s}$ increases as a function of b, the percent of variables from the block that get upweighted. An analytical approximation of this curve (blue) is obtained using Eqn. (7.7), taking the limit $\epsilon \to 0$. The weighting method can do significantly better than naive PCA, shown in black/gray (the predicted value, in green, is lower due to finite n effects). ($p = 3000$, $n = 1000$, $\alpha = 1/3$, $\beta = 10$, $s = 2$, $\epsilon = 0.1$.)

Fig. 7.4 compares PCA performed on X to PCA performed on $\tilde{X}$ (i.e. SCA). The eigenvalue spectra of the two sample covariance matrices are plotted in Fig. 7.4A,B, showing that the largest eigenvalue in SCA increases relative to that of naive PCA. Fig. 7.4D recapitulates the SCA analysis in[57] which used the principal components of $\tilde{X}$ to cluster the variables into three different functionally important protein sectors (colored as red, greed, and blue), or groups of positions that control different protein phenotypes. The clustering, subsequently verified with experimental assays, was performed purely based on the positions of the variables on the biplot. The same clustering methodology applied to the principal components from naive PCA (Fig. 7.4C) would have had a significantly different result, with fewer and different variables clustered in those groups. (Colors in Fig. 7.4C correspond to cluster identities and colors from Fig. 7.4D).

In this example, J and X are not independent, and so SCA might be viewed as a

nonlinear transformation of the data. However, the principal components that result from SCA[28,138] are highly skewed towards those same variables that are upweighted by J. We therefore model SCA using the following simplification: suppose the data is drawn from a normal distribution with covariance matrix $\Sigma = \mathbb{I} + s\mathbf{e_s}\mathbf{e_s}^T$. The components of the signal vector $\mathbf{e_s}$ are normally distributed, but skewed so that the first B entries are larger than the others. Again we multiply $\tilde{X} = JX$, with J a diagonal matrix, where a fraction α of the diagonals are 1 and the rest are $\epsilon < 1$; this downweights some variables relative to others. The choice of J affects the resulting largest eigenvector of $\tilde{X}\tilde{X}^T/n$. If J has 1's for the first B entries and ϵ's for the rest, signal recovery is improved, since the variables which most heavily influence the signal vector are enhanced relative to the unimportant variables (Fig. 7.5A,B). More generally, Fig. 7.5C shows that depending on b, the fraction of the block variables that are upweighted ($0 \leq b \leq 1$), the weighting procedure can either strengthen or weaken signal recovery.

7.5 Discussion

In this work we have explored how PCA is improved when prior information is incorporated, in an effort to formulate a mathematical foundation for pragmatic procedures in the literature. We examined three categories of prior information, differing in the type of information held by the user: knowledge regarding (i) imbalanced variables, (ii) imbalanced samples, and (iii) signal direction. In each case we show that PCA can be markedly improved by employing simple linear transformations incorporating this knowledge. In the most extreme case, a signal may go from being undetectable in the untransformed data to detectable in the transformed data. These models help explain why the many examples of pragmatic PCA in the literature, which typically lack justification, often result in much better signal recovery than that obtained through standard

PCA.

Certainly there are other types of prior information not included in these categories. For example, specialized PCA algorithms which recover low-rank signal have been developed for example to exploit knowledge of presumed sparsity in the principal components (Sparse PCA[158]), or knowledge of sparse corruption of the low-rank data (Robust PCA[21]). It would be interesting to analyze these algorithms from a random matrix theory perspective as well.

It is worth remarking that in principle there is a perhaps more natural method for incorporating prior information into the analysis of covariance data, namely by computing a Bayesian estimate for the covariance matrix by explicitly taking into account the prior information in the Bayesian posterior. The principal eigenvectors of the estimated covariance matrix could then serve as a low-rank representation of the signal. Computation of Bayesian estimates for the covariance matrix may be computationally challenging with domain specific priors, and so Bayesian incorporation of prior information for low-rank reconstruction of signal is not frequently employed. However, there are certain tractable cases. For example, one common prior for the covariance matrix is the Inverse-Wishart distribution. The distribution, whose parameters are a fixed Σ_0 prior covariance matrix and a real-valued number ν, enjoys wide use because it simplifies computation, since it is the conjugate prior for the covariance matrix of normally distributed variables. The resulting posterior for the covariance matrix is thus itself also an Inverse-Wishart distribution, whose mean is given by $\Sigma^* = \frac{XX^T+\Sigma_0}{n+\nu-p-1}$. In general, the matrices XX^T and Σ_0 are not random with respect to each other, as X contains signal which Σ_0 purports to be related to, so it is challenging to use RMT to describe how the eigenvalues and eigenvectors of XX^T differ from those of Σ^*. However, RMT may be used to examine certain simple situations, for example the null model (when there

is no signal in X) to calculate how much the eigenvalues and eigenvectors are affected by Σ_0. The general formulation for examining this uses free probability[109].

The literature has shown a tendency to lean towards 'pragmatic' procedures, often without the statistical rigor of Bayesian or other well-developed methods. Here we have shown that some of these approaches can be understood in the light of prior information incorporation, using random matrix theory to quantify the theoretical limits of such procedures in a rigorous fashion.

8

Outlook

In this dissertation, I have detailed the research I carried out quantitatively analyzing structures in biology, ranging from novel structure discovery to algorithmic developments to new theoretical tools. I would like to conclude with some thoughts about the future directions of the three areas I have covered: genome structure, self-assembly, and signal recovery in noisy datasets. Broadly speaking, the next steps in all three of these fields will be a deeper understanding of the science, and a finer control of the engineering.

In our work on the 3D genome, we uncovered a series of novel structural motifs in the folded genome. In particular, our work provides the first reliable, genome-wide atlas of genomic loops, structures long thought to exist but whose existence, location,

and function remained obscured for decades. These loops are held together by loop anchors which bind convergently oriented CTCF. They correlate with gene activation, and bundle together chromatin into subunits called contact domains, each a contiguous stretch of DNA which self-associates, has a similar pattern of epigenetic marks, and has similar long-range contact patterns.

This work demonstrates definite links between genome folding and fundamental concepts in cellular biology — it suggests that the three-dimensional conformation of the genome is tightly coupled to epigenomics, gene regulation, and cell function. However, there still remain gaping holes in our understanding of these links. The biological process by which a genome folds is still mysterious: how does the genome move to where it needs to go in the nucleus? How do proteins navigate to their appropriate locations along the genome? Frequently, explanations of these processes resort to the stopgap phrase that biological entities are 'recruited' to their appropriate locations in the nucleus. The use of this term is one indication of our deep ignorance of how these processes fundamentally work. For example, the spatial compartmentalization of the genome is highly correlated with histone modifications along the chromatin: chromatin which is modified with activation marks is spatially clustered away from chromatin which is modified with repressive marks. Remarkably, even the direction of causality of this correlation is unknown: are histones first modified, and this then somehow signals where the chromatin should be spatially located, or does all chromatin which resides in a nuclear certain neighborhood get similarly coated with the same histone modifications? The former option is more appealing for many reasons, yet a mechanism by which this could be accomplished is not known. On the other hand, while we are far from understanding the mechanistic link between the 1D features (i.e. epigenetics) and the 3D features (i.e. spatial locations) of the genome, rapid progress is on the horizon for the prediction of

one from the other. Prediction is often far easier than understanding, and the high correlation between 1D and 3D features in the genome makes inference endeavors seem highly feasible.

We found that loops in mammalian genomes are marked by convergently oriented CTCF-bearing anchors. This was a highly unexpected discovery, because it gave an incredibly strict limitation to where loops could form in the genome. Even so, the mechanism of loop formation has not yet been discovered. We hypothesized the existence of the extrusion complex[123], which lands on the chromatin and pulls or extrudes a loop through itself, and only halts the extrusion process when a pair of convergently oriented CTCF motifs is reached. While this model is very consistent with the data, such a complex has yet to be identified in vivo. And though we have found them to be crucial for successful loop formation, the actual role of the CTCF and cohesin proteins in looping remains elusive.

Furthermore, the connection between chromatin looping and gene regulation, which we study in our work, is still enigmatic. We annotated on the order of 10,000 highly enriched specific interactions in the human genome, while previous estimates of classic enhancer-promoter loops put this number at 1,000,000. About a third of the loops we found are involved in connecting a promoter and enhancer, and these are associated with upregulation of the gene at the promoter. How do enhancers accomplish such regulation? They are posited to 'recruit' crucial proteins such as polymerase to increase the transcription of the gene – yet how this is actually accomplished is unclear. More perplexing still is that we find that the majority of active enhancers and promoters are actually not involved in specific looping. If the two loci are not looped together, how can one affect the transcription of the other? One hypothesis is that two loci bundled together in the same contact domain have frequent enough contact with each other that

an enhancer can affect the transcription of a gene even without having a specific interaction with its promoter. Characterizing the connection between enhancers, transcription, and spatial genomics is a crucial direction of this field. In tandem, understanding what happens when the genome is misfolded is an important piece of the story. Studies which identify disease associated genetic variants routinely detect implicative mutations not only in protein coding genes, which is intuitive, but also in noncoding regions of the genome. Mutations which disrupt the ability of a loop to form, for example by mutating the CTCF binding site at a loop anchor, and thereby alter the proper spatial packaging of a gene with other active genomic elements, is one conspicuous mechanism which can explain some of this behavior. Cataloguing the presence of loop disruption in cells which are malfunctioning in some way (e.g. cancer cells) is on the near horizon and provides exciting possibilities for better understanding the genetics of disease.

Our work studying the DXZ4 locus on the inactive X demonstrates tantalizing connections between this tandem CTCF array and X-inactivation. We found similar spatial structures centered at DXZ4 orthologs in the X chromosomes of rhesus macaques and mice. Deletion of DXZ4 on the inactive X resulted in large-scales departures from the normal packaging of the chromosome. And yet, despite its presumably causal role in the inactive X-specific structural features, DXZ4 deletion from the inactive X does not appear to significantly alter the actual transcription of genes from the inactive X (or rather the general lack thereof). Cracking the puzzle of the particular role of DXZ4, the unique 3D conformation of the inactive X, and the silencing of genes remains an exciting future challenge in the field.

We have demonstrated an ability not only to measure the 3D genome structure, but also to modify it. We have shown that deletions or disruptions of CTCF motif sites at looping anchors can lead to striking and predictable changes in the local domain and

looping structure[123]. Similarly, our work on the inactive X here shows that a 100kb deletion of the tandem array of CTCF sites at the DXZ4 locus leads to chromosome wide rearrangements in the 3D structure. Innovations which facilitate the ability to scale-up these types of engineering efforts will lead to spectacular advances in 3D genomics; robust genome-wide editing capabilities would allow us to precisely dictate by design the complete 3D structure of a genome.

The field of self-assembly includes, but is far broader than, genomic structures in the nucleus. The dazzling array of biological structures that assemble through bottom-up processes has inspired an entire field of experimental study. In the past fifteen years, there have been a proliferation of experimental techniques aimed at similarly engineering self-assembled structures, such as via DNA origami or shape complementarity. Yet the complexity of structures built through these techniques has been limited, and central questions in the field have remained unanswered: what sets the complexity cap of these structures? How important of a role does non-equilibrium physics play? How do specific interactions built from vastly different types of physical interactions compare with each other? Our work here improves our understanding of these types of systems by demonstrating that any system composed of specific interactions has a fundamental complexity limit built in from an information theory perspective. This allows us to calculate the capacity, the maximal amount of information that can be encoded using a system of specific interactions and still be resolved by the interactions, as a function of experimentally tunable parameters. Further and more detailed characterization of the physics of experimental systems using this mutual information framework can yield important insights for improving synthetic self-assembly. It has been a longstanding goal to be able to design complex 3D structures which assemble from the bottom up; reaching this goal would be revolutionary, as it will allow for design on the nano and

microscale that is simply not achievable through traditional top-down engineering. As more theory is developed to understand the structures that are achievable through specific interactions, this dream comes closer and closer to reality.

Finally, the study of signal detection in noisy datasets has had a long and illustrious history as a discipline. Yet some of the tools for covariance matrix estimation from random matrix theory are still quite recent, and so it is rare to see them used in data science applications. In many instances of pragmatic PCA in the literature, wherein the data is transformed in some application specific manner to incorporate implicit prior knowledge about the system, very little mathematical justification is given. Our work here demonstrates that mathematical interpretations of certain methods is indeed feasible, and in fact can provide deep intuition as to the efficacy of these algorithms. While our approach does not cover all of the many examples of pragmatic PCA found in the literature, we hope that with careful work further extensions can be possible. Such theoretical results will be useful for calculating the limitations of domain specific datasets and methods, as our work here does for a few examples. This research direction can further be developed as an engineering tool, to aid with the more general problem of tailoring algorithmic methods so that they incorporate highly domain specific prior information.

What is life? Erwin Schrödinger[130] and others have posited that an organism is an entity which, by non-equilibrium processes, consumes free energy from its surroundings to locally decrease its disorder. This fundamental property of life, the local reduction in entropy, results in the creation of ordered structures. While there is assuredly no lack of stochasticity and randomness in biology, the essence of biology is the organization which emerges from this disorder. These structures are unlike many other well-studied materials, and can neither be described by something as orderly as regular crystals nor

as disorderly as random collections of particles. Their very existence brings a wealth of questions to the scientific forefront: What are these structures? What functions do they serve? What sets the limits of how complex these structures can become? How are they assembled in the cell? How can we build similar structures? How can they be measured? The study of life has been the study of biological structures. This thesis is a summary of my contributions to this endeavor.

‘

A

Appendix for Chapter 5

A.1 Mutual Information in Random Ensembles

A.1.1 Derivation

Below we derive Eqn 2. in the main text. Our model system has N distinct "locks" x_1, x_2,..., $x_N \in X$, which each have unique binding partners, "keys" y_1, y_2,..., $y_N \in Y$. We do not specify the physics of these interacting components but require that there is a well defined binding energy $E_{ij} \equiv E(x_i, y_j)$ between every lock and key, and that locks do not bind to locks (and respectively keys to keys). We assume that each lock binds with its cognate key with a strong on-target energy $E_{ii} = s$, while each off target lock and key bind with a weak off-target energy $E_{i,j\neq i} = w \geq s$ which is drawn from

some random distribution. We can equivalently rewrite formulae in terms of $\Delta = w - s$, where $\Delta \sim \rho(\Delta)$.

In this model we assume that there are equal concentrations of locks and keys in a well-mixed solution, and binding probabilities are determined by the Boltzmann distribution such that lock x_i and key y_j bind with probability $p(x_i, y_j) = e^{-\beta E_{ij}}/Z$. Here

$$\begin{aligned} Z &= \sum_{x_i \in X, y_j \in Y} e^{-\beta E_{ij}} && \text{(A.1)} \\ &= \sum_{x_i} \left(e^{-\beta s} + \sum_{y_j, j \neq i} e^{-\beta w_{ij}} \right) && \text{(A.2)} \\ &= e^{-\beta s} \sum_{x_i} \left(1 + \sum_{j \neq i} e^{-\beta \Delta_{ij}} \right) && \text{(A.3)} \end{aligned}$$

As the N pairs of locks and keys are drawn randomly from the ensemble, we replace Z with $\langle Z \rangle$, where the angled brackets denote the average with respect to $\rho(\Delta)$. This gives

$$\langle Z \rangle = e^{-\beta s} N \left[1 + (N-1) \langle e^{-\beta \Delta} \rangle \right] \qquad \text{(A.4)}$$

$p(x_i)$ is the marginal distribution of x_i, representing the total probability of seeing x_i in a bound pair. We approximate

$$\begin{aligned} p(x_i) &= \frac{e^{-\beta s}}{Z} \left(1 + \sum_{j \neq i} e^{-\beta \Delta_{ij}} \right) && \text{(A.5)} \\ &\approx \frac{e^{-\beta s}}{\langle Z \rangle} \left(1 + (N-1) \langle e^{-\beta \Delta} \rangle \right) && \text{(A.6)} \\ &= \tfrac{1}{N} && \text{(A.7)} \end{aligned}$$

In the same manner, $p(y_j) = \frac{1}{N}$.

Mutual information between components X and Y is defined as

$$I(X;Y) = \sum_{x_i \in X, y_j \in Y} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)} \tag{A.8}$$

For a mixture of N pairs of locks and keys, we can rewrite the mutual information as a function of N and the distribution $\rho(\Delta)$.

$$I(N) = \sum_{x_i} \left(\frac{1}{N\left[1 + (N-1)\langle e^{-\beta\Delta}\rangle\right]} \log \frac{N}{1 + (N-1)\langle e^{-\beta\Delta}\rangle} + \sum_{y_j, j \neq i} \frac{e^{-\beta\Delta_{ij}}}{N\left[1 + (N-1)\langle e^{-\beta\Delta}\rangle\right]} \log \frac{N e^{-\beta\Delta_{ij}}}{1 + (N-1)\langle e^{-\beta\Delta}\rangle} \right) \tag{A.9}$$

Replacing the last sum with its expected value, we get

$$\begin{aligned} I(N) &= \log \frac{N}{1 + (N-1)\langle e^{-\beta\Delta}\rangle} - \frac{(N-1)\langle \beta\Delta e^{-\beta\Delta}\rangle}{1 + (N-1)\langle e^{-\beta\Delta}\rangle} && \text{(A.10)} \\ &= \log N_{eff} && \text{(A.11)} \\ N_{eff}(N) &= \frac{N}{1 + (N-1)\langle e^{-\beta\Delta}\rangle} e^{-\frac{(N-1)\langle \beta\Delta e^{-\beta\Delta}\rangle}{1 + (N-1)\langle e^{-\beta\Delta}\rangle}} && \text{(A.12)} \end{aligned}$$

If the log in the above equation is $\log_2$, I will have units of bits.

A.1.2 Two Types of Behavior for I

Examination of Eqn. A.10 reveals two distinct behaviors of the mutual information which depend on $\rho(\Delta)$, and specifically on $\langle \beta\Delta e^{-\beta\Delta}\rangle \geq 0$ and $0 \leq \langle e^{-\beta\Delta}\rangle \leq 1$. When $D \equiv \langle \beta\Delta e^{-\beta\Delta}\rangle - \langle e^{-\beta\Delta}\rangle + \langle e^{-\beta\Delta}\rangle^2 > 0$, I will have a distinct maximum.

Type I - $D > 0$: If interactions have reasonable specificity, typical values of the gap will be at least $1 - 2k_BT$. In this case, the distribution $\rho(\Delta)$ will be mostly supported at $\Delta > 1 - 2k_BT$ and for any such reasonable distribution, $\langle \beta\Delta e^{-\beta\Delta}\rangle > \langle e^{-\beta\Delta}\rangle > \langle e^{-\beta\Delta}\rangle^2$ and hence $D > 0$.

When $D > 0$, we can calculate the capacity as the maximum value of I, occurring at $N = N_C$, as:

$$N_C = \frac{(1 + \langle e^{-\beta\Delta}\rangle)^2}{\langle \beta\Delta e^{-\beta\Delta}\rangle - \langle e^{-\beta\Delta}\rangle + \langle e^{-\beta\Delta}\rangle^2} \tag{A.13}$$

$$C_{D>0} = -\log\langle \beta\Delta e^{-\beta\Delta}\rangle + \frac{\langle \beta\Delta e^{-\beta\Delta}\rangle}{1 - \langle e^{-\beta\Delta}\rangle} + \log(1 - \langle e^{-\beta\Delta}\rangle) - 1 \tag{A.14}$$

$$\approx -\log\langle \beta\Delta e^{-\beta\Delta}\rangle - 1 \tag{A.15}$$

(where for simplicity, C throughout this section is written in units of nats, although it can be converted to the equations in the main text by multiplying by $\log_2(e)$.)

Type II - $D < 0$: For completeness, we discuss the case of $D = \langle \beta\Delta e^{-\beta\Delta}\rangle - \langle e^{-\beta\Delta}\rangle + \langle e^{-\beta\Delta}\rangle^2 < 0$ which can occur for distributions $\rho(\Delta)$ whose typical values are near $\beta\Delta < 1 k_B T$, i.e., typical off-target interactions are as large as on-target binding.

When $D < 0$, either $I > 0$ reaches its maximum at $N = 1$, or at $N = \infty$. In the latter case, we can compute $C = I(N \to \infty)$. Taking the limit of I as N goes to infinity gives

$$C_{D<0} = -\log\langle e^{-\beta\Delta}\rangle - \frac{\langle \beta\Delta e^{-\beta\Delta}\rangle}{\langle e^{-\beta\Delta}\rangle} \tag{A.16}$$

$$\approx -\log\langle e^{-\beta\Delta}\rangle \tag{A.17}$$

(where the last approximation results from $D < 0 \Rightarrow \frac{\langle \beta\Delta e^{-\beta\Delta}\rangle}{\langle e^{-\beta\Delta}\rangle} < 1$).

It is instructive to see these two behaviors for two simple distributions. If the gap

distribution comes from a dirac delta function, $\rho(\Delta) = \delta(\Delta - \mu)$, we find that

$$\begin{aligned} \langle e^{-\beta\Delta} \rangle &= e^{-\beta\mu} && \text{(A.18)} \\ \langle \beta\Delta e^{-\beta\Delta} \rangle &= \beta\mu e^{-\beta\mu} && \text{(A.19)} \\ D &= e^{-\beta\mu}(\beta\mu + e^{-\beta\mu} - 1) && \text{(A.20)} \\ &> 0 && \text{(A.21)} \end{aligned}$$

Thus the mutual information behaves as Type I, and

$$\begin{aligned} N_C &= \frac{e^{\beta\mu} - 2 + e^{-\beta\mu}}{\beta\mu - 1 + e^{-\beta\mu}} && \text{(A.22)} \\ &\to e^{\beta\mu}\frac{1}{\beta\mu} && \text{(A.23)} \\ C &= \log(\frac{e^{\beta\mu} - 1}{\beta\mu}) + \frac{\beta\mu}{e^{\beta\mu} - 1} - 1 && \text{(A.24)} \\ &\to \beta\mu && \text{(A.25)} \end{aligned}$$

as $\beta\mu \to \infty$. Gap energy distributions $\rho(\Delta)$ which are tightly centered at a large mean value (gaussian, poisson, etc.) behave similarly.

On the other hand, if Δ is drawn from an exponential distribution, $\rho(\Delta) = \frac{1}{\mu}e^{-\Delta/\mu}$, the behavior is different. In this case:

$$\begin{aligned} \langle e^{-\beta\Delta} \rangle &= \frac{1}{1 + \beta\mu} && \text{(A.26)} \\ \langle \beta\Delta e^{-\beta\Delta} \rangle &= \frac{\beta\mu}{(1 + \beta\mu)^2} && \text{(A.27)} \\ D &= 0 && \text{(A.28)} \end{aligned}$$

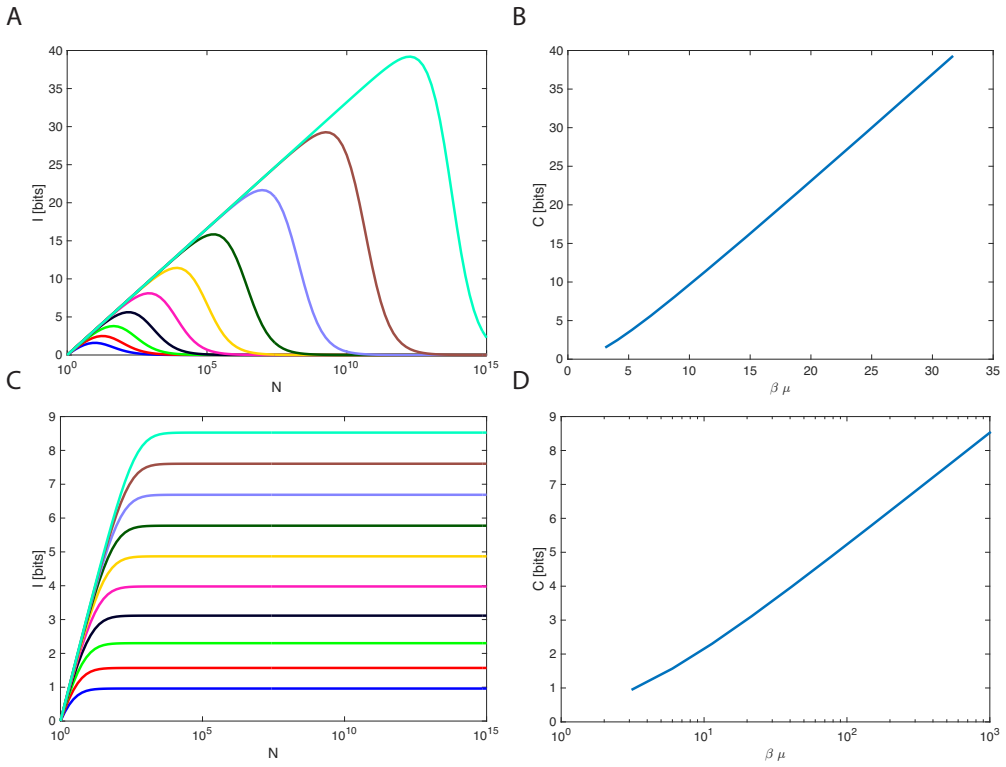


Figure A.1: A) Mutual information plotted for the delta function distribution $\rho(\Delta) = \delta(\Delta - \mu)$ for various (log spaced) $\beta\mu$. I displays characteristic rise with $\log(N)$, maximum, and then a decay at large N. **B)** The capacity is plotted for $\rho(\Delta) = \delta(\Delta - \mu)$. As $\beta\mu$ gets large, capacity is linear. **C)** Mutual information plotted for the exponential function distribution $\rho(\Delta) = \frac{1}{\mu}e^{-\Delta/\mu}$ for various (log spaced) $\beta\mu$. I displays a characteristic rise with $\log(N)$, and then plateaus at its maximum. **D)** The capacity as a function of $\beta\mu$ for the exponential distribution. As $\beta\mu$ gets large, the capacity goes as $\log(\beta\mu)$ (the x-axis is log scale).

The mutual information for a system with this gap distribution thus does not exhibit a maximum, but rather follows the equation for Type II.

$$C = \log(1+\beta\mu) - \frac{\beta\mu}{1+\beta\mu} \tag{A.29}$$

$$\rightarrow \log(\beta\mu) \tag{A.30}$$

as $\beta\mu \rightarrow \infty$. (For small $\beta\mu \ll 1$, we can expand the log to get $C \approx (1-\beta\mu)^2 \sim (\beta\mu)^2$). This is exactly what we see when we look at the mutual information numerically.

A.1.3 Normalization Choice

To compute mutual information, one needs to first transform a matrix of binding energies between N locks and N keys into one of probabilities. Above we have chosen $p(x_i, y_j) = e^{-\beta E_{ij}}/Z$, where $Z = \sum_{x_i \in X, y_j \in Y} e^{-\beta E_{ij}}$. This corresponds to a model where all the locks and keys are placed in a well-mixed enclosure together and every key has the chance to bind to every lock.

An alternative model might choose a different normalization. For example, to encompass analyte sensing in synthetic applications and protein sensors in biological signal transduction pathways, a slightly different model, which would also more closely reflect the Shannon communication setting, is necessary. For these systems one may choose a model in which the locks (which here we think of as sensors) are held in place at a given location, while each key (the object to be detected) enters, and has a chance to bind to each of the locks. In this case, the probability of a key y_j should sum to one over all possible locks. Thus $p(x_i, y_j) = e^{-\beta E_{ij}}/Z_j$, and $Z_j = \sum_{x_i \in X} e^{-\beta E_{ij}}$.

While results for the two normalizations closely resemble one another, a significant difference can arise for example when the kth lock/key pair that does not bind very

well ($|E_{kk}|$ is small), but is highly specific ($|E_{kk}| \gg |E_{k,j\neq k}|$). In this case, the kth pair would not contribute to the mutual information in the well-mixed case (since it rarely binds), while in the communication setting, it would contribute.

A.1.4 Relationship between $\hat{\Delta}$, $\bar{s}$, $\bar{w}$

In the text we have defined $\beta\hat{\Delta} = -\log\langle\beta\Delta e^{-\beta\Delta}\rangle$, where the average is taken over the probability distribution $\rho(\Delta)$ and $\Delta = w - s > 0$. If s is the same for all lock-key pairs (as is the case, for example, with DNA binding), we may rewrite this as

$$\begin{aligned}
\beta\hat{\Delta} &= -\log\langle\beta(w-s)e^{-\beta(w-s)}\rangle && \text{(A.31)}\\
&= -\log(e^{\beta s}) - \log\left(-\beta s\langle e^{-\beta w}\rangle + \langle\beta w e^{-\beta w}\rangle\right) && \text{(A.32)}\\
&= -\beta\bar{s} - \log\left(-\beta s\langle e^{-\beta w}\rangle + \langle\beta w e^{-\beta w}\rangle\right) && \text{(A.33)}\\
&\approx -\beta\bar{s} - \log\langle e^{-\beta w}\rangle - \log(-\beta s) && \text{(A.34)}\\
&= \beta\bar{w} - \beta\bar{s} - \log(-\beta s) && \text{(A.35)}
\end{aligned}$$

where the approximation follows for $|s| \gg |w|$.

A.1.5 Numerical Sampling Methods for Random Ensembles

To numerically calculate the capacity of a random ensemble of lock/key pairs, two options are available. First, for each N, one can randomly sample many NxN matrices, drawn from the ensemble, and calculate the average $I(N)$ at that point using Eqn. A.8 above. The capacity C is then given by the maximum I achieved, and N_C just the N at which I reaches C.

Alternatively, one can use Eqn. A.10 above. This requires calculating $\langle e^{-\beta\Delta}\rangle$ and $\langle\beta\Delta e^{-\beta\Delta}\rangle$. If an analytical formula for these terms is available, this may be computed

directly. However, even in complex energy scenarios, as long as one can randomly sample pairs of locks and keys and measure their binding energy, one can calculate these terms numerically and then use Eqn. A.10. These two give the same results in most cases.

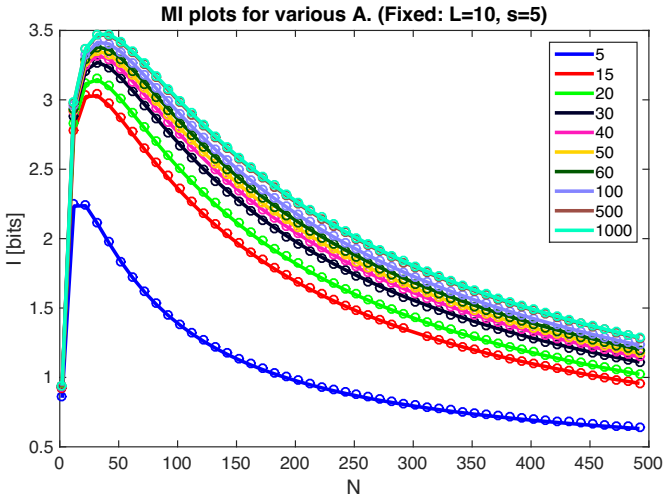


Figure A.2: Averaging over NxN matrices and 2x2 matrices gives the same result. We simulate our color model with $L = 10$, $s = -5$, $\beta = 1$, for various A's marked in the legend. Lines: NxN simulation, circles: 2x2 sampling.

A.2 Case study: Binomial binding

It is instructive to see how capacity scales for the case of binomially distributed gap, which serves as a model for binding in color systems. Here we assume that each on target pair of letters bind with energy $-\epsilon\ k_BT$, while an off target pair of letters bind with energy 0 k_BT. (One may also allow for a nonzero binding energy between off-target letters - and in fact, we do so for Figure 5 C, for which on-target letters bind with $\epsilon_{x^l=y^l} = 0.5\ k_BT$ and off-target letters bind with $\epsilon_{x^l\neq y^l} = 0.25\ k_BT$. However, for simplicity, the rest of the paper and the analysis below sets off-target letter binding to zero.)

If crosstalk is negligible ($\beta\epsilon = \infty$), then we can use all A^L possible pairs, and $C = L\log_2 A$. However, when crosstalk plays a significant role, we must use Eqn A.10 above to determine the capacity and $N_C < A^L$. Plugging in $\rho(\Delta) = \frac{1}{\epsilon} f_{binom}(\Delta/\epsilon, L, 1 - 1/A)$ for the distribution, we can calculate the capacity using the approximation from Eqn. 5 (main text) as:

$$C \approx \beta\epsilon L \log_2(e) - \log_2(e\beta\epsilon L) + L\log_2 \frac{A}{A - 1 + e^{\beta\epsilon}} + \log_2\left(1 + \frac{e^{\beta\epsilon}}{(A-1)}\right) \quad \text{(A.36)}$$

One interesting consequence of this is that $C(2L) > 2C(L)$. In other words, more information can be encoded if a single channel is composed of a longer sequence than if many channels of smaller sequences are used. However, the difference is logarithmic in L and so is negligible for large L.

We may examine how the capacity scales as A, the alphabet size, becomes large.

$$\lim_{A\to\infty} C = \beta\epsilon L\log_2(e) - \log_2(e\beta\epsilon L) \quad \text{(A.37)}$$

$$\lim_{A\to\infty} N_c = \frac{(e^{\beta\epsilon L} - 1)^2}{1 + e^{\beta\epsilon L}(\beta\epsilon L - 1)} \quad \text{(A.38)}$$

From above one can see that $\beta\epsilon$ is what sets the upper limit of the capacity in the crosstalk-limited regime. Since on-target binding is fixed as $s = -\epsilon L$, capacity is determined by w. In the worst case scenario, an on-target lock looks identical to an off-target lock, which would mean every letter binds, giving $w_{ij}^{worst} = s$, and thus a gap (given by $\Delta = w - s$) of $\Delta^{worst} = 0$. This would occur if $A = 1$, in other words when the surfaces are just sticky. On the other hand, in the best case scenario, none of the letters in x_i bind to the letters y_j, giving $w_{ij}^{best} = 0$ and $\Delta^{best} = -s = \epsilon L$. This occurs when $A \to \infty$. Thus A is just a tuning factor that allows the system to go from w^{worst} to w^{best}

with increasing A: as A is increased, the more likely that a letter in x_i is mismatched to a letter in y_j, giving a zero contribution for that pair.

Furthermore the weak binding is given by

$$\bar{w} = -L \log \frac{A - 1 + e^{\beta\epsilon}}{A} \tag{A.39}$$

$$\lim_{L\to\infty} \bar{w}(\epsilon = s/L) = -\beta s/A \tag{A.40}$$

The mutual information in the binomial case has a maximum at $N_C < \infty$, but plateaus down to a constant for large N. We can derive that limit by taking the limit of I as $N \to \infty$ of Eqn. A.10. Alternatively, we can derive the limit from Eqn. A.8 without any averaging, by assuming that when $N \gg A^L$, the total possible unique pairs, the information content should be the same as using each of the A^L pairs exactly once. In either case we find:

$$I(N \to \infty) = \beta\epsilon L \log_2(e) + L \log_2 \frac{A}{A - 1 + e^{\beta\epsilon}} \tag{A.41}$$

A.3 Fabrication Defects

Unintended defects in synthesis can cause mismatch between cognate pairs and thus reduce on-target binding and thus capacity. We model such defects by adding independent (normally distributed) random undulations of order σ to keys which would otherwise be cognate to their lock, thereby reducing on-target binding. We find that the effect of such defects is small when $\sigma \ll d$, the depletion particle size, but degrades capacity significantly when $\sigma \sim d$. Such degradation originates from the same aspect of shape coding that gives it lower crosstalk than color coding; a single large mismatch can be sufficient to dramatically reduce on-target binding.

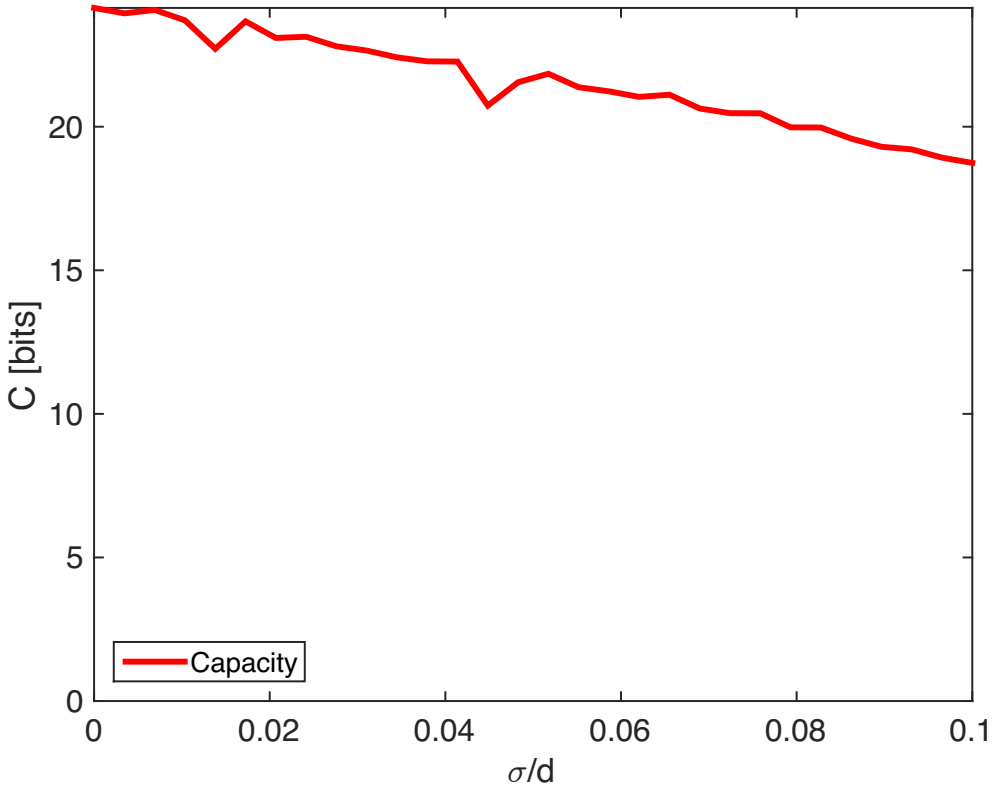


Figure A.3: Fabrication defects of order σ degrade the capacity when $\sigma \sim d$. In the main text we assume $\sigma \ll d$. Here $L = 10$ and $d = 0.2\mu$m.

A.4 Pac-man Particle Interactions

We derive the depletion-driven interactions between pairs of spherical lock/key Pac-man particles of the type described in (14). We describe each lock (Fig A.4A) as a sphere containing a hemispherical cavity of radius r_l, cut off at angle θ (though all calculations in the main text were performed for $\theta = \pi/2$). Each key is a sphere of size r_k, and for a cognate lock and key pair, $r_k = r_l$. Attraction is mediated by depletant particles of diameter $d \ll r_k, r_l$, and the energy of attraction between a lock and key is $E = -\epsilon V_{\text{exc}}$, where V_{exc} is the change of excluded depletant particle volume (and ϵ mediates the

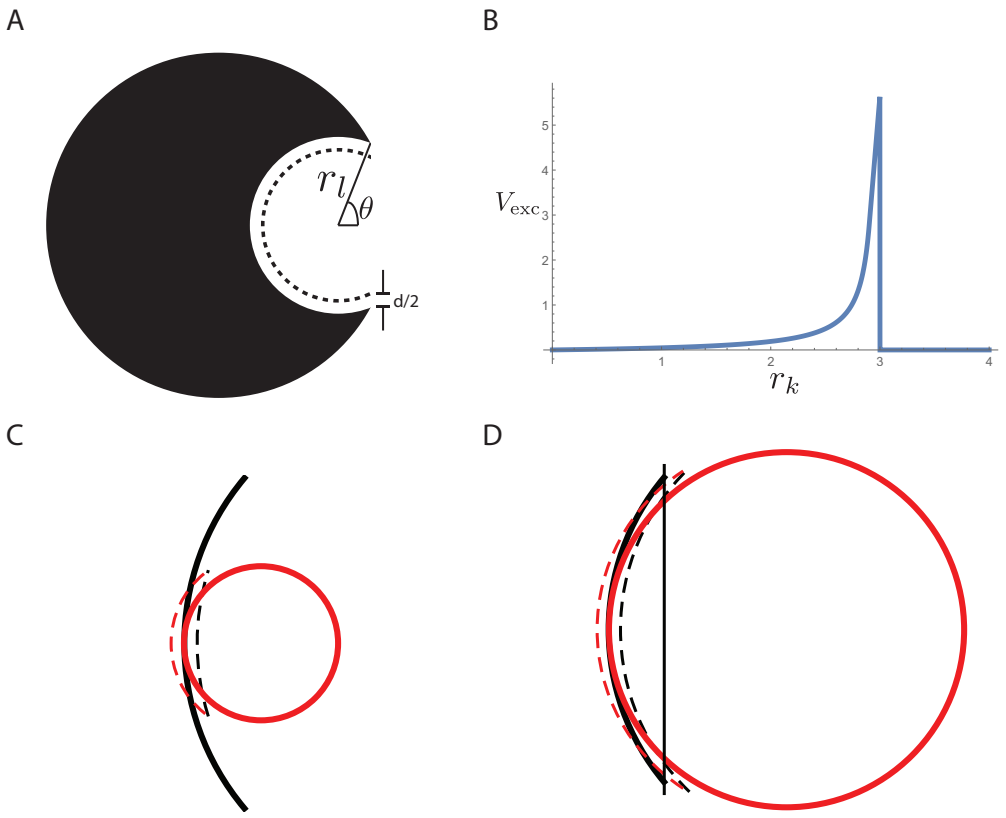


Figure A.4: (A) Model lock. Without key binding, depletant particles are excluded from a volume $d\Omega r_l^2$ in the cavity of the lock. **(B)** Full calculation for the change in excluded volume as a function of r_k. Here $r_l = 3$ and $\theta = \pi/2$. **(C)** When $r_k \ll r_l$, the excluded volume is just given by the contact area of the two spheres. **(D)** When r_k approaches r_l, the intersection with the plane must be taken into account.

strength of the attraction). We derive a formula for V_{exc} that is more accurate than that used in (14), particularly when the radii of the lock and key are close to each other (eg. $r_l - r_k = O(d)$).

We derived V_{exc} (Fig A.4B) for four regimes of lock/key sizes:

When $\underline{r_k = r_l}$, $V_{\text{exc}} = dA_{\text{contact}}$, where A_{contact} is the contact area between the lock and the key. The contact area is simply the entire inner area of the lock, $A_{\text{contact}} = \Omega r_l^2 = 2\pi(1 - \cos\theta)r_l^2$ where θ is half the total angle subtended by the inner surface of

the lock at its origin.

When $r_k > r_l$, the key does not fit into the lock and makes contact along an annulus of radius $2\pi r_l \cos\theta$. As discussed in (14), the width t of this annulus is set by fabrication constraints (in particular, how sharp the edges of the locks are) and was estimated to be $t = 190$ nm. Hence the effective contact area is $A_{\text{contact}} = t(2\pi r_l \cos\theta)$ and $V_{\text{exc}} = dA_{\text{contact}}$. We assume that the contact area drops to this value at $r_k = r_l + 2d$ from the value of Ωr_l^2 at $r_k = r_l$, and linearly interpolate between the two.

When $r_k < r_l$, we consider two regimes:

(1) If $r_k \ll r_l$: In this case, we wish to compute the excluded volume shown as the area between dotted lines in the Fig A.4C. The excluded volume is entirely contained within the lock; in particular, the edges of the lock play no role. Hence, the excluded volume is easily computed as the intersection of two overlapping spheres. We find,

$$V_{\text{exc}} = \pi d^2 \left(d/6 - \frac{r_k r_l}{r_k - r_l} \right) \tag{A.42}$$

(2) When r_k approaches r_l from below, at a particular radius $r_{\text{thresh}} < r_l$, the excluded volume reaches the edges of the lock. See Fig A.4D; using the formula for intersecting spheres would give an overestimate of the binding energy. Instead, we draw a plane through the opening of the lock (line in Fig A.4D) and only count the excluded volume contained between the plane and the lock. The radius r_{thresh} at which such edge effects become important is given by,

$$r_{\text{thresh}} = \frac{r_l(d/2 - r_l)(\cos\theta - 1)}{d/2 + r_l + (d/2 - r_l)\cos\theta} \tag{A.43}$$

The formula for a triple intersection for $r_{\text{thresh}} < r_k < r_l$ gives,

$$
\begin{aligned}
x &= r_l \cos\theta + d^2/(12 r_l) && \text{(A.44)} \\
h_k &= r_l + d/2 - x && \text{(A.45)} \\
v_k &= h_k^2 \pi((r_k + d/2) - h_k/3) && \text{(A.46)} \\
h_l &= r_l - d/2 - x && \text{(A.47)} \\
v_l &= h_l^2 \pi((rl - d/2) - h_l/3) && \text{(A.48)} \\
V_{\text{exc}} &= vk - vl && \text{(A.49)}
\end{aligned}
$$

The full curve for V_{exc} is shown in Fig A.4B.

A.5 Contact Between Random Surfaces

In the main text, we emphasized that shape-based coding achieves better specificity than color-based coding because the amount of crosstalk for shapes is much smaller than for color. The intuitive idea is that when coding with shapes, off-target binding between two mismatched random shapes typically constitutes only a few points of contact and does not scale linearly with the size of the components. We verified this by simulating off-target pairs of shaped components and measuring the average $|w|$, the off-target binding energy, as a function of L and d. As shown in Fig. A.5A, the binding energy first rises sublinearly with L and then saturates for large L. When we allow translations between the components (Fig. A.5B), such that w is taken as the strongest possible off-target binding across all translations, the curve is still strongly sublinear in L (indeed, highly logarithmic).

We may further verify this by examining random 1-d walks of length L, and measuring

how much time they typically spend within a distance d of their maximum. We simulate random walks $x(t)$, where in each 'time' step, the random walk can take a step of $x(t+1) = x(t) \pm 1$. Examining a range of d from 0.5 to 5 of various lengths, we find that the average amount of time spent within d of the maximum first increases sublinearly with L, and then flattens out to a constant (Fig. A.5C). Two typical walks of $L = 300$ and $L = 20,000$, at $d = 3$, give intuition for why this is the case (Fig. A.5D): when L is small relative to d, a good fraction of the walk is contained between d of its maximum, and so increasing L increases the amount of time spent in this area.

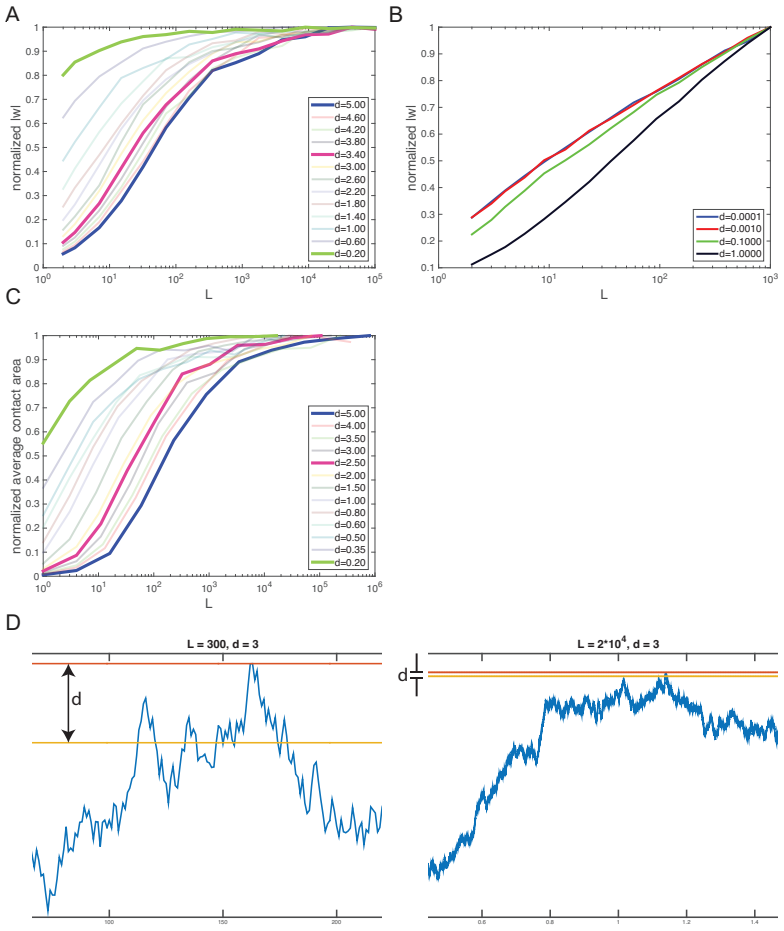


Figure A.5: A) We measure the average off target energy $|w|$ for shaped components (of the form shown in Fig. 3A, main text) at various d and L. No translation is allowed between components. Each curve is normalized by its maximal value. As L increases, $|w|$ first increases sublinearly with L, and then saturates. **B)** The normalized $|w|$ for shaped components when translation is allowed - w for each pair is taken as the strongest possible binding across all translations. Each curve is normalized by its maximal value. As L increases, $|w|$ increases logarithmically with L. There does not appear to be saturation for the range of numerically achievable L. **C)** Similarly, we measure the average amount of time a random 1-d walk of L 'time' steps spends within d of its maximum. As L becomes large with respect to d, the average time saturates. Each curve is normalized by its maximal value. **D)** Two typical random walks of length 300 (left) and 20,000 (right). The amount of time that the random walk spends between its maximum (red line) and within $d = 3$ of its maximum (yellow line) is a large fraction of the length for small L, but for very long walks is independent of L.

References

[1] Adleman, L. M. (1994). Molecular computation of solutions to combinatorial problems. *Science*, 266(5187), 1021–1024.

[2] Ahmadiyeh, N., Pomerantz, M. M., Grisanzio, C., Herman, P., Jia, L., Almendro, V., He, H. H., Brown, M., Liu, X. S., Davis, M., Caswell, J. L., Beckwith, C. A., Hills, A., MacConaill, L., Coetzee, G. A., Regan, M. M., & Freedman, M. L. (2010). 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with myc. *Proceedings of the National Academy of Sciences*, 107(21), 9742–9746. 10.1073/pnas.0910668107.

[3] Alipour, E. & Marko, J. F. (2012). Self-organization of domain structures by dna-loop-extruding enzymes. *Nucleic Acids Res*, 40(22), 11202–12.

[4] Alivisatos, A. P., Johnsson, K. P., Peng, X., Wilson, T. E., Loweth, C. J., Bruchez, M. P., & Schultz, P. G. (1996). Organization of'nanocrystal molecules' using DNA. 382, 609–611.

[5] Alon, N. (1998). The shannon capacity of a union. *Combinatorica*, 18(3), 301–310.

[6] Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H., & Shiroishi, T. (2009). Chromosomal dynamics at the shh locus: limb bud-specific differential regulation of competence and active transcription. *Developmental cell*, 16(1), 47–57. 10.1016/j.devcel.2008.11.011.

[7] Bai, Z. & Silverstein, J. W. (2009). Spectral Analysis of Large Dimensional Random Matrices. *Springer Series in Statistics*, (pp. 1–560).

[8] Baik, J., Arous, G. B., & Péché, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5), pp. 1643–1697.

[9] Banerji, J., Rusconi, S., & Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote sv40 dna sequences. *Cell*, 27(2 Pt 1), 299–308.

[10] Barr, M. & Bertram, E. (1949). A morphological distinction between neurones of the male and female, and the behaviour of the nucleolar satellite during accelerated nucleoprotein synthesis. *Nature*, 163, 676–677.

[11] Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D., Wang, Z., Wei, G., Chepelev, I., & Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4), 823–837.

[12] Beliveau, B., Joyce, E., Apostolopoulos, N., Yilmaz, F., Fonseka, C., McCole, R., Chang, Y., Li, J., Senaratne, T., Williams, B., Rouillard, J.-M., & Wu, C.-t. (2012). Versatile design and synthesis platform for visualizing genomes with oligopaint fish probes. *Proceedings of the National Academy of Sciences of the United States of America*, 109(52), 21301–21306. 10.1073/pnas.1213818110.

[13] Benaych-Georges, F. & Nadakuditi, R. R. (2011). The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1), 494 – 521.

[14] Berletch, J. B., Ma, W., Yang, F., Shendure, J., Noble, W. S., Disteche, C. M., & Deng, X. (2015). Escape from x inactivation varies in mouse tissues. *PLoS Genet*, 11(3), e1005079.

[15] Biancaniello, P., Kim, A., & Crocker, J. (2005). Colloidal Interactions and Self-Assembly Using DNA Hybridization. *Physical Review Letters*, 94(5), 058302–058307.

[16] Bickmore, W. (2013). The spatial organization of the human genome. *Annual review of genomics and human genetics*, 14, 67–84. 10.1146/annurev-genom-091212-153515.

[17] Blackwood, E. & Kadonaga, J. (1998). Going the distance: a current view of enhancer action. *Science (New York, N.Y.)*, 281(5373), 60–63. 10.1126/science.281.5373.60.

[18] Bloemendal, A., Knowles, A., Yau, H.-T., & Yin, J. (2014). On the principal components of sample covariance matrices. *ArXiv e-prints*.

[19] Bloemendal, A. & Virág, B. (2013). Limits of spiked random matrices i. *Probability Theory and Related Fields*, 156(3-4), 795–825.

[20] Bourque, G., Leong, B., Vega, V. B., Chen, X., Lee, Y. L., Srinivasan, K. G., Chew, J.-L. L., Ruan, Y., Wei, C.-L. L., Ng, H. H., & Liu, E. T. (2008). Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome research*, 18(11), 1752–1762. 10.1101/gr.080663.108.

[21] Candès, E. J., Li, X., Ma, Y., & Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3), 11.

[22] Carrel, L. & Willard, H. F. (2005). X-inactivation profile reveals extensive variability in x-linked gene expression in females. *Nature*, 434(7031), 400–4.

[23] Carrillo, R. A., Özkan, E., Menon, K. P., Nagarkar-Jaiswal, S., Lee, P.-T., Jeon, M., Birnbaum, M. E., Bellen, H. J., Garcia, K. C., & Zinn, K. (2015). Control of synaptic connectivity by a network of drosophila igsf cell surface proteins. *Cell*, 163(7), 1770–1782.

[24] Chadwick, B. (2008). Dxz4 chromatin adopts an opposing conformation to that of the surrounding chromosome and acquires a novel inactive x-specific role involving ctcf and antisense transcripts. *Genome research*, 18(8), 1259–1269. 10.1101/gr.075713.107.

[25] Chadwick, B. P. (2007). Variation in xi chromatin organization and correlation of the h3k27me3 chromatin territories to transcribed sequences by microarray analysis. *Chromosoma*, 116(2), 147–57.

[26] Chadwick, B. P. & Willard, H. F. (2004). Multiple spatially distinct types of facultative heterochromatin on the human inactive x chromosome. *Proc Natl Acad Sci U S A*, 101(50), 17450–5.

[27] Chandola, V., Banerjee, A., & Kumar, V. (2007). Outlier detection: A survey. *ACM Computing Surveys*.

[28] Colwell, L. J., Brenner, M. P., & Murray, A. W. (2014a). Conservation weighting functions enable covariance analyses to detect functionally important amino acids. *PloS one*, 9(11), e107723.

[29] Colwell, L. J., Qin, Y., Huntley, M., Manta, A., & Brenner, M. P. (2014b). Feynman-hellmann theorem and signal identification from sample covariance matrices. *Physical Review X*, 4(3), 031032.

[30] Consortium, E. P., Bernstein, B., Birney, E., Dunham, I., Green, E., Gunter, C., & Snyder, M. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414), 57–74. 10.1038/nature11247.

[31] Cook, P. R. & Brazell, I. A. (1975). Supercoils in human dna. *Journal of cell science*, 19(2), 261–279.

[32] Cover, T. M. & Thomas, J. A. (2012). *Elements of information theory*. John Wiley & Sons.

[33] Cremer, M., von Hase, J., Volm, T., Brero, A., Kreth, G., Walter, J., Fischer, C., Solovei, I., Cremer, C., & Cremer, T. (2001). Non-random radial higher-order chromatin arrangements in nuclei of diploid human cells. *Chromosome research*, 9(7), 541–567.

[34] Cremer, T. & Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Reviews: Genetics*, 2(4), 292–301.

[35] Cremer, T., Cremer, C., Baumann, H., Luedtke, E., Sperling, K., Teuber, V., & Zorn, C. (1982). Rabl's model of the interphase chromosome arrangement tested in chinise hamster cells by premature chromosome condensation and laser-uv-microbeam experiments. *Human genetics*, 60(1), 46–56.

[36] Cuddapah, S., Jothi, R., Schones, D., Roh, T.-Y., Cui, K., & Zhao, K. (2009). Global analysis of the insulator binding protein ctcf in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome research*, 19(1), 24–32. 10.1101/gr.082800.108.

[37] Cullen, K., Kladde, M., & Seyfred, M. (1993). Interaction between transcription regulatory regions of prolactin chromatin. *Science*, 261(5118), 203–206. 10.1126/science.8327891.

[38] Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing chromosome conformation. *Science*, 295(5558), 1306–1311. 10.1126/science.1067799.

[39] Deng, X., Ma, W., Ramani, V., Hill, A., Yang, F., Ay, F., Berletch, J. B., Blau, C. A., Shendure, J., Duan, Z., Noble, W. S., & Disteche, C. M. (2015). Bipartite structure of the inactive mouse x chromosome. *Genome Biol*, 16, 152.

[40] Denkov, N., Tcholakova, S., Lesov, I., Cholakova, D., & Smoukov, S. K. (2015). Self-shaping of oil droplets via the formation of intermediate rotator phases upon cooling. *Nature*, 528(7582), 392–395.

[41] Dixon, J., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398), 376–380. 10.1038/nature11082.

[42] Dostie, J., Richmond, T., Arnaout, R., Selzer, R., Lee, W., Honan, T., Rubio, E., Krumm, A., Lamb, J., Nusbaum, C., Green, R., & Dekker, J. (2006). Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. *Genome research*, 16(10), 1299–1309. 10.1101/gr.5571506.

[43] El Karoui, N. (2005). Recent results about the largest eigenvalue of random covariance matrices and statistical application. *Acta Physica Polonica Series B*, 36(9), 2681.

[44] Finlan, L. E., Sproul, D., Thomson, I., Boyle, S., Kerr, E., Perry, P., Ylstra, B., Chubb, J. R., & Bickmore, W. A. (2008). Recruitment to the nuclear periphery can alter expression of genes in human cells. *PLoS genetics*, 4(3). 10.1371/journal.pgen.1000039.

[45] Fischer, E. (1894). Einfluss der configuration auf die wirkung der enzyme. *Berichte der deutschen chemischen Gesellschaft*, 27(3), 2985–2993.

[46] Fromm, M. & Berg, P. (1983). Simian virus 40 early-and late-region promoter functions are enhanced by the 72-base-pair repeat inserted at distant locations and inverted orientations. *Molecular and cellular biology*, 3(6), 991–999.

[47] Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., & Mirny, L. A. (2015). Formation of chromosomal domains by loop extrusion. *bioRxiv*.

[48] Fullwood, M., Liu, M., Pan, Y., Liu, J., Xu, H., Mohamed, Y., Orlov, Y., Velkov, S., Ho, A., Mei, P., Chew, E., Huang, P., Welboren, W.-J., Han, Y., Ooi, H., Ariyaratne, P., Vega, V., Luo, Y., Tan, P., Choy, P., Wansa, K., Zhao, B., Lim, K., Leow, S., Yow, J., Joseph, R., Li, H., Desai, K., Thomsen, J., Lee, Y., Karuturi, R., Herve, T., Bourque, G., Stunnenberg, H., Ruan, X., Cacheux-Rataboul, V.,

Sung, W.-K., Liu, E., Wei, C.-L., Cheung, E., & Ruan, Y. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, 462(7269), 58–64. 10.1038/nature08497.

[49] Gaszner, M. & Felsenfeld, G. (2006). Insulators: exploiting transcriptional and epigenetic mechanisms. *Nature Reviews: Genetics*, 7(9), 703–713. 10.1038/nrg1925.

[50] Gavrilov, A. A., Gushchanskaya, E. S., Strelkova, O., Zhironkina, O., Kireev, I. I., Iarovaia, O. V., & Razin, S. V. (2013). Disclosure of a structural milieu for the proximity ligation reveals the elusive nature of an active chromatin hub. *Nucleic acids research*, 41(6), 3563–3575. 10.1093/nar/gkt067.

[51] Giacalone, J., Friedes, J., & Francke, U. (1992). A novel gc-rich human macrosatellite vntr in xq24 is differentially methylated on active and inactive x chromosomes. *Nat. Genet.*, 1(2), 137–143.

[52] Gil, A. M., David, M. A., Richard, M. D., Gonçalo, R. A., David, R. B., Aravinda, C., Andrew, G. C., Peter, D., Evan, E. E., Paul, F., Stacey, B. G., Richard, A. G., Eric, D. G., Matthew, E. H., Bartha, M. K., Jan, O. K., Eric, S. L., Charles, L., Hans, L., Elaine, R. M., Gabor, T. M., Gil, A. M., Deborah, A. N., Jeanette, P. S., Stephen, T. S., Jun, W., Richard, K. W., Richard, A. G., Huyen, D., Christie, K., Sandra, L., Lora, L., Donna, M., Jeff, R., Min, W., Jun, W., Xiaodong, F., Xiaosen, G., Min, J., Hui, J., Xin, J., Guoqing, L., Jingxiang, L., Yingrui, L., Zhuo, L., Xiao, L., Yao, L., Xuedi, M., Zhe, S., Shuaishuai, T., Meifang, T., Bo, W., Guangbiao, W., Honglong, W., Renhua, W., Ye, Y., Wenwei, Z., Jiao, Z., Meiru, Z., Xiaole, Z., Yan, Z., Eric, S. L., David, M. A., Stacey, B. G., Namrata, G., Paul, F., Laura, C., Rasko, L., Richard, E. S., Xiangqun, Z.-B., David, R. B., Russell, G., Sean, H., Terena, J., Zoya, K., Hans, L., Ralf, S., Marcus, W. A., Vyacheslav, S. A., Tatiana, A. B., Matthias, L., Florian, M., Marc, S., Bernd, T., Marie-Laure, Y., Stephen, T. S., Gil, A. M., Elaine, R. M., Richard, K. W., Lucinda, F., Robert, F., George, M. W., Richard, M. D., Senduran, B., John, B., Petr, D., Thomas, M. K., Anja, K.-K., Shane, M., James, S., et al. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491.

[53] Gilad, Y. & Mizrahi-Man, O. (2015). A reanalysis of mouse encode comparative gene expression data [version 1; referees: 3 approved, 1 approved with reservations]. *F1000Research*, 4(121).

[54] Goldman, M. (1988). The chromatin domain as a unit of gene regulation. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 9(2-3), 50–55.

[55] Gracias, D. H., Tien, J., Breen, T. L., Hsu, C., & Whitesides, G. M. (2000). Forming electrical networks in three dimensions by self-assembly. *Science*, 289(5482), 1170–1172.

[56] Hahn, M. A., Wu, X., Li, A. X., Hahn, T., & Pfeifer, G. P. (2011). Relationship between gene body dna methylation and intragenic h3k9me3 and h3k36me3 chromatin marks. *PLoS ONE*.

[57] Halabi, N., Rivoire, O., Leibler, S., & Ranganathan, R. (2009). Protein sectors: evolutionary units of three-dimensional structure. *Cell*, 138(4), 774–786.

[58] Hansen, R. S., Thomas, S., Sandstrom, R., Canfield, T. K., Thurman, R. E., Weaver, M., Dorschner, M. O., Gartler, S. M., & Stamatoyannopoulos, J. A. (2010). Sequencing newly replicated dna reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences of the United States of America*, 107(1), 139–144.

[59] Hedges, L. O., Mannige, R. V., & Whitelam, S. (2014). Growth of equilibrium structures built from a large number of distinct component types . *Soft Matter*, 10(34), 6404–6416.

[60] Heidari, N., Phanstiel, D. H., He, C., Grubert, F., Jahanbanian, F., Kasowski, M., Zhang, M. Q., & Snyder, M. P. (2014). Genome-wide map of regulatory interactions in the human genome. *Genome research*.

[61] Hernandez, C. J. & Mason, T. G. (2007). Colloidal alphabet soup: monodisperse dispersions of shape-designed lithoparticles. *The Journal of Physical Chemistry C*, 111(12), 4477–4480.

[62] Hoffman, M. M., Ernst, J., Wilder, S. P., Kundaje, A., Harris, R. S., Libbrecht, M., Giardine, B., Ellenbogen, P. M., Bilmes, J. A., Birney, E., Hardison, R. C.,

Dunham, I., Kellis, M., & Noble, W. S. (2012). Integrative annotation of chromatin elements from encode data. *Nucleic acids research*, 41(2), 827–841.

[63] Horakova, A. H., Calabrese, J. M., McLaughlin, C. R., Tremblay, D. C., Magnuson, T., & Chadwick, B. P. (2012). The mouse dxz4 homolog retains ctcf binding and proximity to pls3 despite substantial organizational differences compared to the primate macrosatellite. *Genome Biol*, 13(8), R70.

[64] Hou, C., Zhao, H., Tanimoto, K., & Dean, A. (2008). Ctcf-dependent enhancer-blocking by alternative chromatin loop formation. *Proceedings of the National Academy of Sciences of the United States of America*, 105(51), 20398–20403. 10.1073/pnas.0808506106.

[65] Huntley, M. H., Murugan, A., & Brenner, M. P. (2016). Information capacity of specific interactions. *Proceedings of the National Academy of Sciences*, (pp. 201520969).

[66] Imakaev, M., Fudenberg, G., McCord, R., Naumova, N., Goloborodko, A., Lajoie, B., Dekker, J., & Mirny, L. (2012). Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nature methods*, 9(10), 999–1003.

[67] Itzkovitz, S., Tlusty, T., & Alon, U. (2006). Coding limits on the number of transcription factors. *BMC genomics*, 7(1), 239.

[68] Jacobs, W. M., Reinhardt, A., & Frenkel, D. (2015). Communication: Theoretical prediction of free-energy landscapes for complex self-assembly. *The Journal of chemical physics*, 142(2), 021101.

[69] Jin, F., Li, Y., Dixon, J., Selvaraj, S., Ye, Z., Lee, A., Yen, C.-A., Schmitt, A., Espinoza, C., & Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, 503(7475), 290–294. 10.1038/nature12644.

[70] Johnson, M. E. & Hummer, G. (2011). Nonspecific binding limits the number of proteins in a cell and shapes their interaction networks. *Proc. Natl. Acad. Sci. U.S.A.*, 108(2), 603–608.

[71] Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, 29(2), pp. 295–327.

[72] Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., & Chen, L. (2012). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nature biotechnology*, 30(1), 90–98. 10.1038/nbt.2057.

[73] Kane, R. S. (2010). Thermodynamics of multivalent interactions: influence of the linker. *Langmuir*, 26(11), 8636–8640.

[74] Ke, Y., Ong, L. L., Shih, W. M., & Yin, P. (2012). Three-Dimensional Structures Self-Assembled from DNA Bricks. *Science*, 338(6111), 1177–1183.

[75] King, N. P., Sheffler, W., Sawaya, M. R., Vollmar, B. S., Sumida, J. P., Andre, I., Gonen, T., Yeates, T. O., & Baker, D. (2012). Computational Design of Self-Assembling Protein Nanomaterials with Atomic Level Accuracy. *Science*, 336(6085), 1171–1174.

[76] Knight, P. & Ruiz, D. (2012). A fast algorithm for matrix balancing. *IMA Journal of Numerical Analysis*. 10.1093/imanum/drs019.

[77] Kornberg, R. D. (1974). Chromatin structure: a repeating unit of histones and dna. *Science*, 184(4139), 868–871.

[78] Küpper, K., Kölbl, A., Biener, D., Dittrich, S., von Hase, J., Thormeyer, T., Fiegler, H., Carter, N. P., Speicher, M. R., Cremer, T., et al. (2007). Radial chromatin positioning is shaped by local gene density, not by gene expression. *Chromosoma*, 116(3), 285–306.

[79] Kurukuti, S., Tiwari, V. K., Tavoosidana, G., Pugacheva, E., Murrell, A., Zhao, Z., Lobanenkov, V., Reik, W., & Ohlsson, R. (2006). Ctcf binding at the h19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to igf2. *Proceedings of the National Academy of Sciences*, 103(28), 10684–10689.

[80] Lai, Y.-T., King, N. P., & Yeates, T. O. (2012). Principles for designing ordered protein assemblies. *Trends in Cell Biology*.

[81] Li, G., Ruan, X., Auerbach, R., Sandhu, K., Zheng, M., Wang, P., Poh, H., Goh, Y., Lim, J., Zhang, J., Sim, H., Peh, S., Mulawadi, F., Ong, C., Orlov, Y., Hong, S., Zhang, Z., Landt, S., Raha, D., Euskirchen, G., Wei, C.-L., Ge, W.,

Wang, H., Davis, C., Fisher-Aylor, K., Mortazavi, A., Gerstein, M., Gingeras, T., Wold, B., Sun, Y., Fullwood, M., Cheung, E., Liu, E., Sung, W.-K., Snyder, M., & Ruan, Y. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 148(1-2), 84–98. 10.1016/j.cell.2011.12.014.

[82] Li, H. & Durbin, R. (2010). Fast and accurate long-read alignment with burrows-wheeler transform. *Bioinformatics (Oxford, England)*, 26(5), 589–595. 10.1093/bioinformatics/btp698.

[83] Lieberman-Aiden, E., van Berkum, N., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B., Sabo, P., Dorschner, M., Sandstrom, R., Bernstein, B., Bender, M., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L., Lander, E., & Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950), 289–293. 10.1126/science.1181369.

[84] Lin, S., Lin, Y., Nery, J. R., Urich, M. A., Breschi, A., Davis, C. A., Dobin, A., Zaleski, C., Beer, M. A., Chapman, W. C., Gingeras, T. R., Ecker, J. R., & Snyder, M. P. (2014). Comparison of the transcriptional landscapes between human and mouse tissues. *Proceedings of the National Academy of Sciences*, 111(48), 17224–17229.

[85] Lingenfelter, P. A., Adler, D. A., Poslinski, D., Thomas, S., Elliott, R. W., Chapman, V. M., & Disteche, C. M. (1998). Escape from x inactivation of smcx is preceded by silencing during mouse development. *Nat Genet*, 18(3), 212–3.

[86] Lockless, S. W. & Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438), 295–299.

[87] Marčenko, V. A. & Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Sbornik: Mathematics*, 1(4), 457–483.

[88] Mason, T. G. (2002). Osmotically driven shape-dependent colloidal separations. *Phys. Rev. E*, 66(6), 060402.

[89] Mason, T. G. (2015). Process for creating shape-designed particles in a fluid. US Patent Office.

[90] McCord, R., Nazario-Toole, A., Zhang, H., Chines, P., Zhan, Y., Erdos, M., Collins, F., Dekker, J., & Cao, K. (2013). Correlated alterations in genome organization, histone methylation, and dna-lamin a/c interactions in hutchinson-gilford progeria syndrome. *Genome research*, 23(2), 260–269.

[91] McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation dna sequencing data. *Genome research*, 20(9), 1297–1303.

[92] McLaughlin, C. R. & Chadwick, B. P. (2011). Characterization of dxz4 conservation in primates implies important functional roles for ctcf binding, array expression and tandem repeat organization on the x chromosome. *Genome Biology*, 12, R37.

[93] Milenkovic, O. & Kashyap, N. (2006). On the design of codes for dna computing. In *Coding and Cryptography* (pp. 100–119). Springer.

[94] Minajigi, A., Froberg, J. E., Wei, C., Sunwoo, H., Kesner, B., Colognori, D., Lessing, D., Payer, B., Boukhali, M., Haas, W., & Lee, J. T. (2015). A comprehensive xist interactome reveals cohesin repulsion and an rna-directed chromosome conformation. *Science*.

[95] Mirkin, C. A., Letsinger, R. L., Mucic, R. C., & Storhoff, J. J. (1996). A DNA-based method for rationally assembling nanoparticles into macroscopic materials. *Nature*, 382(6592), 607–609.

[96] Miszta, K., de Graaf, J., Bertoni, G., Dorfs, D., Brescia, R., Marras, S., Ceseracciu, L., Cingolani, R., van Roij, R., Dijkstra, M., & Manna, L. (2011). Hierarchical self-assembly of suspended branched colloidal nanocrystals into superlattice structures. *Nat Mater*, 10(11), 872–876.

[97] Mukherjee, S., Erickson, H., & Bastia, D. (1988). Enhancer-origin interaction in plasmid r6k involves a dna loop mediated by initiator protein. *Cell*, 52(3), 375–383. 10.1016/S0092-8674(88)80030-8.

[98] Murugan, A., Zou, J., & Brenner, M. P. (2015). Undesired usage and the robust self-assembly of heterogeneous structures. *Nature communications*, 6.

[99] Myers, C. R. (2008). Satisfiability, sequence niches and molecular codes in cellular signalling. *IET Systems Biology*, 2(5), 304.

[100] Nadakuditi, R. & Silverstein, J. (2010). Fundamental limit of sample generalized eigenvalue based detection of signals in noise using relatively few signal-bearing and noise-only samples. *Selected Topics in Signal Processing, IEEE Journal of*, 4(3), 468–480.

[101] Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E. D., Tanay, A., & Fraser, P. (2013). Single-cell hi-c reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469), 59–64. 10.1038/nature12593.

[102] Nasmyth, K. (2001). Disseminating the genome: joining, resolving, and separating sister chromatids during mitosis and meiosis. *Annu Rev Genet*, 35, 673–745.

[103] Naughton, C., Avlonitis, N., Corless, S., Prendergast, J., Mati, I., Eijk, P., Cockroft, S., Bradley, M., Ylstra, B., & Gilbert, N. (2013). Transcription forms and remodels supercoiling domains unfolding large-scale chromatin structures. *Nature structural & molecular biology*, 20(3), 387–395.

[104] Naughton, C., Sproul, D., Hamilton, C., & Gilbert, N. (2010). Analysis of active and inactive x chromosome architecture reveals the independent organization of 30 nm and large-scale chromatin structures. *Mol Cell*, 40(3), 397–409.

[105] Nora, E., Lajoie, B., Schulz, E., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J., & Heard, E. (2012). Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, 485(7398), 381–385. 10.1038/nature11049.

[106] Nozawa, R. S., Nagao, K., Igami, K. T., Shibata, S., Shirai, N., Nozaki, N., Sado, T., Kimura, H., & Obuse, C. (2013). Human inactive x chromosome is compacted through a prc2-independent smchd1-hbix1 pathway. *Nat Struct Mol Biol*, 20(5), 566–73.

[107] Nykypanchuk, D., Maye, M. M., van der Lelie, D., & Gang, O. (2008). DNA-guided crystallization of colloidal nanoparticles. *Nature*, 451(7178), 549–552.

[108] Oudet, P., Gross-Bellard, M., & Chambon, P. (1975). Electron microscopic and biochemical evidence that chromatin structure is a repeating unit. *Cell*, 4(4), 281–300.

[109] Pastur, L. & Vasilchuk, V. (2000). On the law of addition of random matrices. *Communications in Mathematical Physics*, 214(2), 249–286.

[110] Pathak, S. (1976). Chromosome banding techniques. *The Journal of reproductive medicine*, 17(1), 25–28.

[111] Péché, S. (2009). Universality results for the largest eigenvalues of some sample covariance matrix ensembles. *Probability Theory and Related Fields*, 143(3), 481–516.

[112] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-lern: Machine learning in python. *Journal of Machine Learning Research*, 12, 2825–2830.

[113] Perelson, A. S. & Oster, G. F. (1979). Theoretical studies of clonal selection: minimal antibody repertoire size and reliability of self-non-self discrimination. *Journal of Theoretical Biology*, 81(4), 645–670.

[114] Pfaffel, O. & Schlemm, E. (2012). Limiting spectral distribution of a new random matrix model with dependence across rows and columns. *Linear Algebra and its Applications*, 436(9), 2966–2979.

[115] Phillips, J. & Corces, V. (2009). Ctcf: master weaver of the genome. *Cell*, 137(7), 1194–1211. 10.1016/j.cell.2009.06.001.

[116] Podgornaia, A. I. & Laub, M. T. (2013). Determinants of specificity in two-component signal transduction. *Current Opinion in Microbiology*, 16(2), 156 – 162.

[117] Podgornaia, A. I. & Laub, M. T. (2015). Pervasive degeneracy and epistasis in a protein-protein interface. *Science*, 347(6222), 673–677.

[118] Randhawa, J. S., Kanu, L. N., Singh, G., & Gracias, D. H. (2010). Importance of surface patterns for defect mitigation in three-dimensional self-assembly. *Langmuir*, 26(15), 12534–12539.

[119] Rao, S. S., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., & Aiden, E. L. (2014). A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7), 1665–80.

[120] Rossi, L., Soni, V., Ashton, D. J., Pine, D. J., Philipse, A. P., Chaikin, P. M., Dijkstra, M., Sacanna, S., & Irvine, W. T. M. (2015). Shape-sensitive crystallization in colloidal superball fluids. *Proceedings of the National Academy of Sciences*, 112(17), 5286–5290.

[121] Ryba, T., Hiratani, I., Lu, J., Itoh, M., Kulik, M., Zhang, J., Schulz, T., Robins, A., Dalton, S., & Gilbert, D. (2010). Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome research*, 20(6), 761–770.

[122] Sacanna, S., Irvine, W., Chaikin, P. M., & Pine, D. J. (2010). Lock and key colloids. *Nature*, 464(7288), 575–578.

[123] Sanborn, A. L., Rao, S. S., Huang, S.-C., Durand, N. C., Huntley, M. H., Jewett, A. I., Bochkov, I. D., Chinnappan, D., Cutkosky, A., Li, J., et al. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences*, 112(47), E6456–E6465.

[124] Sanyal, A., Lajoie, B., Jain, G., & Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature*, 489(7414), 109–113. 10.1038/nature11279.

[125] Savir, Y. & Tlusty, T. (2009). Molecular recognition as an information channel: The role of conformational changes. *Information Sciences and Systems, 2009. CISS 2009. 43rd Annual Conference on*, (pp. 835–840).

[126] Schaap, M., Lemmers, R. J., Maassen, R., van der Vliet, P. J., Hoogerheide, L. F., van Dijk, H. K., Basturk, N., de Knijff, P., & van der Maarel, S. M. (2013).

Genome-wide analysis of macrosatellite repeat copy number variation in worldwide populations: evidence for differences and commonalities in size distributions and size restrictions. *BMC Genomics*, 14, 143.

[127] Schleif, R. (1992). Dna looping. *Annual review of biochemistry*, 61(1), 199–223.

[128] Schmidt, D., Schwalie, P. C., Wilson, M. D., Ballester, B., Gonçalves, ., Kutter, C., Brown, G. D., Marshall, A., Flicek, P., & Odom, D. T. (2012). Waves of retrotransposon expansion remodel genome organization and ctcf binding in multiple mammalian lineages. *Cell*, 148(1-2), 335–348. 10.1016/j.cell.2011.11.058.

[129] Schneider, T. (1997). Information content of individual genetic sequences. *Journal of Theoretical Biology*, 189(4), 427 – 441.

[130] Schrödinger, E. (1992). *What is life?: With mind and matter and autobiographical sketches*. Cambridge University Press.

[131] Schulz, E. G. & Heard, E. (2013). Role and control of x chromosome dosage in mammalian development. *Curr Opin Genet Dev*, 23(2), 109–15.

[132] Segel, L. A. & Perelson, A. S. (1989). Shape space: an approach to the evaluation of cross-reactivity effects, stability and controllability in the immune system. *Immunology letters*, 22(2), 91.

[133] Sexton, T., Schober, H., Fraser, P., & Gasser, S. (2007). Gene regulation through nuclear organization. *Nature structural & molecular biology*, 14(11), 1049–1055. 10.1038/nsmb1324.

[134] Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., & Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the drosophila genome. *Cell*, 148(3), 458–472. 10.1016/j.cell.2012.01.010.

[135] Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., & de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4c). *Nature genetics*, 38(11), 1348–1354.

[136] Splinter, E., Heath, H., Kooren, J., Palstra, R.-J., Klous, P., Grosveld, F., Galjart, N., & de Laat, W. (2006). Ctcf mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Genes & development*, 20(17), 2349–2354. 10.1101/gad.399506.

[137] Teller, K., Illner, D., Thamm, S., Casas-Delucchi, C. S., Versteeg, R., Indemans, M., Cremer, T., & Cremer, M. (2011). A top-down analysis of xa- and xi-territories reveals differences of higher order structure at >/= 20 mb genomic length scales. *Nucleus*, 2(5), 465–77.

[138] Teşileanu, T., Colwell, L. J., & Leibler, S. (2015). Protein sectors: statistical coupling analysis versus conservation. *PLoS Comput Biol*, 11(2), e1004091.

[139] Tolhuis, B., Palstra, R., Splinter, E., Grosveld, F., & de Laat, W. (2002). Looping and interaction between hypersensitive sites in the active beta-globin locus. *Molecular cell*, 10(6), 1453–1465. 10.1016/S1097-2765(02)00781-5.

[140] Tremblay, D. C., Moseley, S., & Chadwick, B. P. (2011). Variation in array size, monomer composition and expression of the macrosatellite dxz4. *PLoS ONE*, 6(4), e18969.

[141] Valignat, M.-P., Theodoly, O., Crocker, J. C., Russel, W. B., & Chaikin, P. M. (2005). Reversible self-assembly and directed assembly of dna-linked micrometer-sized colloids. *Proceedings of the National Academy of Sciences of the United States of America*, 102(12), 4225–4229.

[142] Vogel, M. J., Guelen, L., de Wit, E., & Hupkes, D. P. (2006). Human heterochromatin proteins form large domains containing krab-znf genes. *Genome ….* 10.1101/gr.5391806.

[143] Vogelstein, B., Pardoll, D. M., & Coffey, D. S. (1980). Supercoiled loops and eucaryotic dna replicaton. *Cell*, 22(1 Pt 1), 79–85.

[144] Watson, J. D., Crick, F. H., et al. (1953). Molecular structure of nucleic acids. *Nature*, 171(4356), 737–738.

[145] Winfree, E. (1998). Algorithmic Self-assembly of DNA. California Institute of Technology.

[146] Woolston, C. (2015). Potential flaws in genomics paper scrutinized on twitter. *Nature News*, 521(7553).

[147] Wu, K.-T., Feng, L., Sha, R., Dreyfus, R., Grosberg, A. Y., Seeman, N. C., & Chaikin, P. M. (2012). Polygamous particles. *Proceedings of the National Academy of Sciences*, 109(46), 18731–18736.

[148] Xie, X., Mikkelsen, T., Gnirke, A., Lindblad-Toh, K., Kellis, M., & Lander, E. (2007). Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of ctcf insulator sites. *Proceedings of the National Academy of Sciences of the United States of America*, 104(17), 7145–7150. 10.1073/pnas.0701811104.

[149] Yaffe, E. & Tanay, A. (2011). Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics*, 43(11), 1059–1065. 10.1038/ng.947.

[150] Yang, S.-M., Kim, S.-H., Lim, J.-M., & Yi, G.-R. (2008). Synthesis and assembly of structured colloidal particles. *J. Mater. Chem.*, 18, 2177–2190.

[151] Ye, X., Collins, J. E., Kang, Y., Chen, J., Chen, D. T. N., Yodh, A. G., & Murray, C. B. (2010). Morphologically controlled synthesis of colloidal upconversion nanophosphors and their shape-directed self-assembly. *Proceedings of the National Academy of Sciences*, 107(52), 22430–22435.

[152] Yusufzai, T. M., Tagami, H., Nakatani, Y., & Felsenfeld, G. (2004). Ctcf tethers an insulator to subnuclear sites, suggesting shared insulator mechanisms across species. *Molecular cell*, 13(2), 291–298.

[153] Zehnbauer, B. A. & Vogelstein, B. (1985). Supercoiled loops and the organization of replication and transcription in eukaryotes. *BioEssays*. 10.1002/bies.950020203.

[154] Zeravcic, Z., Manoharan, V. N., & Brenner, M. P. (2014). Size limits of self-assembled colloidal structures made using specific interactions. *Proceedings of the National Academy of Sciences*, 111(45), 15918–15923.

[155] Zhao, K., Bruinsma, R., & Mason, T. G. (2011). Entropic crystal–crystal transitions of brownian squares. *Proceedings of the National Academy of Sciences*, 108(7), 2684–2687.

[156] Zhao, K. & Mason, T. G. (2007). Directing colloidal self-assembly through roughness-controlled depletion attractions. *Physical review letters*, 99(26), 268301.

[157] Zhao, K. & Mason, T. G. (2008). Suppressing and enhancing depletion attractions between surfaces roughened by asperities. *Physical review letters*, 101(14), 148301.

[158] Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2), 265–286.